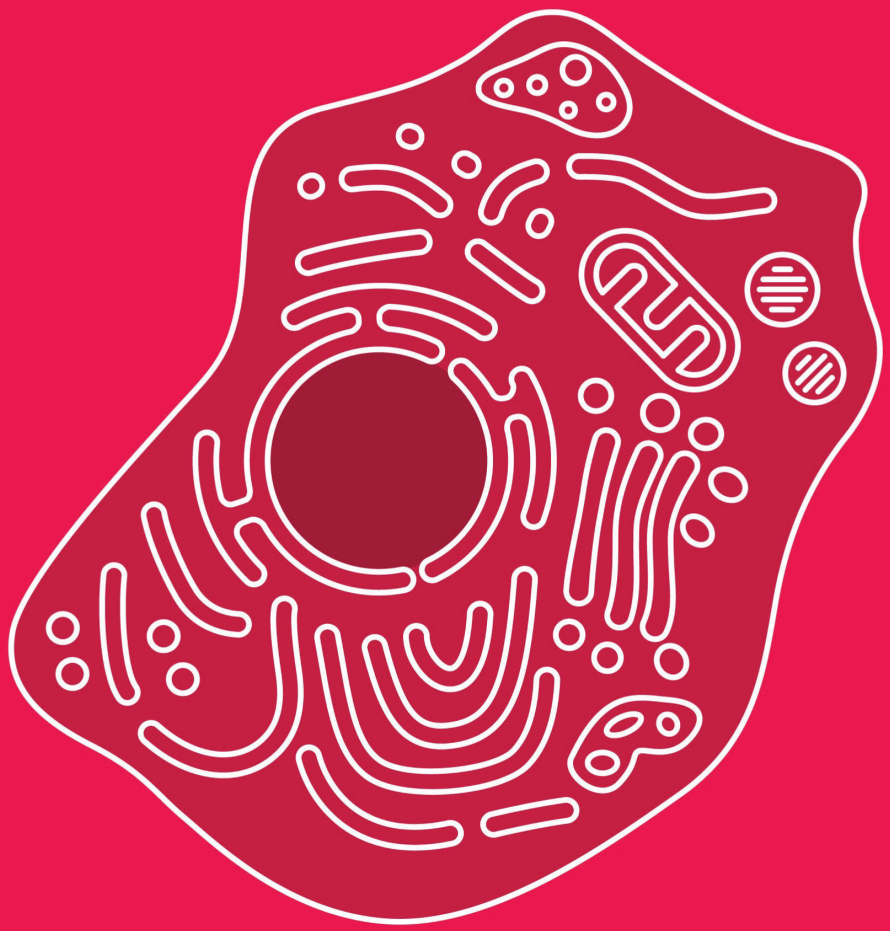
# Molecular Biology of THE CELL

Fifth Edition



ALBERTS JOHNSON LEWIS RAFF ROBERTS WALTER

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 1 of 240

This page intentionally left blank.

# Molecular Biology of THE CELL

Fifth Edition

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 3 of 240

This page intentionally left blank.

# Molecular Biology of THE CELL

Fifth Edition

Bruce Alberts

Alexander Johnson

Julian Lewis

Martin Raff

Keith Roberts

Peter Walter

*With problems by*

John Wilson

Tim Hunt

Garland Science
Taylor & Francis Group

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 5 of 240

*Garland Science*
Vice President: Denise Schanck
Assistant Editor: Sigrid Masson
Production Editor and Layout: Emma Jeffcock
Senior Publisher: Jackie Harbor
Illustrator: Nigel Orme
Designer: Matthew McClements, Blink Studio, Ltd.
Editors: Marjorie Anderson and Sherry Granum
Copy Editor: Bruce Goatly
Indexer: Merrall-Ross International, Ltd.
Permissions Coordinator: Mary Dispenza

*Cell Biology Interactive*
Artistic and Scientific Direction: Peter Walter
Narrated by: Julie Theriot
Production Design and Development: Michael Morales

© 2008, 2002 by Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter.
© 1983, 1989, 1994 by Bruce Alberts, Dennis Bray, Julian Lewis, Martin Raff, Keith Roberts, and James D. Watson.

**Bruce Alberts** received his Ph.D. from Harvard University and is Professor of Biochemistry and Biophysics at the University of California, San Francisco. For 12 years, he served as President of the U.S. National Academy of Sciences (1993–2005). **Alexander Johnson** received his Ph.D. from Harvard University and is Professor of Microbiology and Immunology and Director of the Biochemistry, Cell Biology, Genetics, and Developmental Biology Graduate Program at the University of California, San Francisco. **Julian Lewis** received his D.Phil. from the University of Oxford and is a Principal Scientist at the London Research Institute of Cancer Research UK. **Martin Raff** received his M.D. from McGill University and is at the Medical Research Council Laboratory for Molecular Cell Biology and the Biology Department at University College London. **Keith Roberts** received his Ph.D. from the University of Cambridge and is Emeritus Fellow at the John Innes Centre, Norwich. **Peter Walter** received his Ph.D. from The Rockefeller University in New York and is Professor and Chairman of the Department of Biochemistry and Biophysics at the University of California, San Francisco, and an Investigator of the Howard Hughes Medical Institute.

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. A wide variety of references are listed. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

All rights reserved. No part of this book covered by the copyright heron may be reproduced or used in any format in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, or information storage and retrieval systems—without permission of the publisher.

**Library of Congress Cataloging-in-Publication Data**
Molecular biology of the cell / Bruce Alberts … [et al.].-- 5th ed.
p. cm
ISBN 978-0-8153-4105-5 (hardcover)---ISBN 978-0-8153-4106-2 (paperback)
1. Cytology. 2. Molecular biology. I. Alberts, Bruce.
QH581.2 .M64 2008
571.6--dc22
2007005475 CIP

Published by Garland Science, Taylor & Francis Group, LLC, an informa business, 270 Madison Avenue, New York NY 10016, USA, and 2 Park Square, Milton Park, Abingdon, OX14 4RN, UK.

Printed in the United States of America

15 14 13 12 11 10 9 8 7 6 5 4 3 2

# 3 Proteins

When we look at a cell through a microscope or analyze its electrical or biochemical activity, we are, in essence, observing proteins. Proteins constitute most of a cell's dry mass. They are not only the cell's building blocks; they also execute nearly all the cell's functions. Thus, enzymes provide the intricate molecular surfaces in a cell that promote its many chemical reactions. Proteins embedded in the plasma membrane form channels and pumps that control the passage of small molecules into and out of the cell. Other proteins carry messages from one cell to another, or act as signal integrators that relay sets of signals inward from the plasma membrane to the cell nucleus. Yet others serve as tiny molecular machines with moving parts: *kinesin*, for example, propels organelles through the cytoplasm; *topoisomerase* can untangle knotted DNA molecules. Other specialized proteins act as antibodies, toxins, hormones, antifreeze molecules, elastic fibers, ropes, or sources of luminescence. Before we can hope to understand how genes work, how muscles contract, how nerves conduct electricity, how embryos develop, or how our bodies function, we must attain a deep understanding of proteins.

## In This Chapter

THE SHAPE AND STRUCTURE OF PROTEINS 125

PROTEIN FUNCTION 152

## THE SHAPE AND STRUCTURE OF PROTEINS

From a chemical point of view, proteins are by far the most structurally complex and functionally sophisticated molecules known. This is perhaps not surprising, once we realize that the structure and chemistry of each protein has been developed and fine-tuned over billions of years of evolutionary history. Yet, even to experts, the remarkable versatility of proteins can seem truly amazing.

In this section, we consider how the location of each amino acid in the long string of amino acids that forms a protein determines its three-dimensional shape. Later in the chapter, we use this understanding of protein structure at the atomic level to describe how the precise shape of each protein molecule determines its function in a cell.

### The Shape of a Protein Is Specified by Its Amino Acid Sequence

There are 20 types of amino acids in proteins, each with different chemical properties. A **protein** molecule is made from a long chain of these amino acids, each linked to its neighbor through a covalent peptide bond. Proteins are therefore also known as *polypeptides*. Each type of protein has a unique sequence of amino acids, and there are many thousands of different proteins, each with its own particular amino acid sequence.

The repeating sequence of atoms along the core of the polypeptide chain is referred to as the **polypeptide backbone**. Attached to this repetitive chain are those portions of the amino acids that are not involved in making a peptide bond and that give each amino acid its unique properties: the 20 different amino acid **side chains** (**Figure 3–1**). Some of these side chains are nonpolar and hydrophobic ("water-fearing"), others are negatively or positively charged, some readily form covalent bonds, and so on. **Panel 3–1** (pp. 128–129) shows their atomic structures and **Figure 3–2** lists their abbreviations.



LCY Biotechnology Holding, Inc.
Ex. 1008
Page 7 of 240

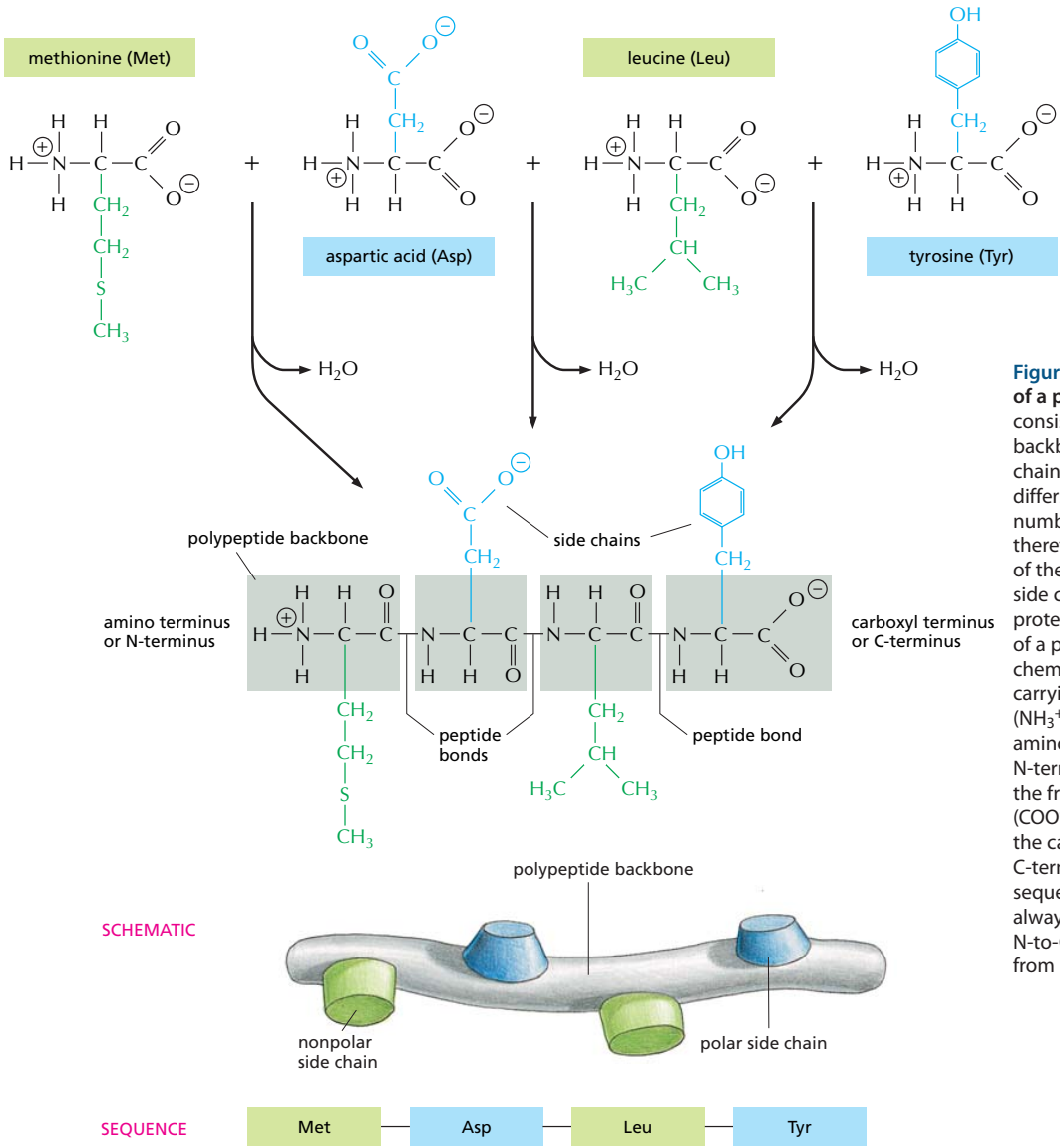



**Figure 3–1 The components of a protein.** A protein consists of a polypeptide backbone with attached side chains. Each type of protein differs in its sequence and number of amino acids; therefore, it is the sequence of the chemically different side chains that makes each protein distinct. The two ends of a polypeptide chain are chemically different: the end carrying the free amino group ($NH_3^+$, also written $NH_2$) is the amino terminus, or N-terminus, and that carrying the free carboxyl group ($COO^-$, also written COOH) is the carboxyl terminus or C-terminus. The amino acid sequence of a protein is always presented in the N-to-C direction, reading from left to right.

As discussed in Chapter 2, atoms behave almost as if they were hard spheres with a definite radius (their *van der Waals radius*). The requirement that no two atoms overlap limits greatly the possible bond angles in a polypeptide chain (**Figure 3–3**). This constraint and other steric interactions severely restrict the possible three-dimensional arrangements of atoms (or *conformations)*. Nevertheless, a long flexible chain, such as a protein, can still fold in an enormous number of ways.

The folding of a protein chain is, however, further constrained by many different sets of weak *noncovalent bonds* that form between one part of the chain and another. These involve atoms in the polypeptide backbone, as well as atoms in the amino acid side chains. There are three types of weak bonds: *hydrogen bonds*, *electrostatic attractions*, and *van der Waals attractions*, as explained in Chapter 2 (see p. 54). Individual noncovalent bonds are 30–300 times weaker than the typical covalent bonds that create biological molecules. But many weak bonds acting in parallel can hold two regions of a polypeptide chain tightly together. In this way, the combined strength of large numbers of such noncovalent bonds determines the stability of each folded shape (**Figure 3–4**).

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 8 of 240

| AMINO ACID | | | SIDE CHAIN |
|---|---|---|---|
| Aspartic acid | Asp | D | negative |
| Glutamic acid | Glu | E | negative |
| Arginine | Arg | R | positive |
| Lysine | Lys | K | positive |
| Histidine | His | H | positive |
| Asparagine | Asn | N | uncharged polar |
| Glutamine | Gln | Q | uncharged polar |
| Serine | Ser | S | uncharged polar |
| Threonine | Thr | T | uncharged polar |
| Tyrosine | Tyr | Y | uncharged polar |

POLAR AMINO ACIDS

| AMINO ACID | | | SIDE CHAIN |
|---|---|---|---|
| Alanine | Ala | A | nonpolar |
| Glycine | Gly | G | nonpolar |
| Valine | Val | V | nonpolar |
| Leucine | Leu | L | nonpolar |
| Isoleucine | Ile | I | nonpolar |
| Proline | Pro | P | nonpolar |
| Phenylalanine | Phe | F | nonpolar |
| Methionine | Met | M | nonpolar |
| Tryptophan | Trp | W | nonpolar |
| Cysteine | Cys | C | nonpolar |

NONPOLAR AMINO ACIDS

**Figure 3–2 The 20 amino acids found in proteins.** Each amino acid has a three-letter and a one-letter abbreviation. There are equal numbers of polar and nonpolar side chains; however, some side chains listed here as polar are large enough to have some non-polar properties (for example, Tyr, Thr, Arg, Lys). For atomic structures, see Panel 3–1 (pp. 128–129).

A fourth weak force also has a central role in determining the shape of a protein. As described in Chapter 2, hydrophobic molecules, including the nonpolar side chains of particular amino acids, tend to be forced together in an aqueous environment in order to minimize their disruptive effect on the hydrogen-bonded network of water molecules (see p. 54 and Panel 2–2, pp. 108–109). Therefore, an important factor governing the folding of any protein is the distribution of its polar and nonpolar amino acids. The nonpolar (hydrophobic) side chains in a protein—belonging to such amino acids as phenylalanine, leucine, valine, and tryptophan—tend to cluster in the interior of the molecule (just as hydrophobic oil droplets coalesce in water to form one large droplet). This enables them to avoid contact with the water that surrounds them inside a cell. In contrast, polar groups—such as those belonging to arginine, glutamine, and histidine—tend to arrange themselves near the outside of the molecule, where they can form hydrogen bonds with water and with other polar molecules (**Figure 3–5**). Polar amino acids buried within the protein are usually hydrogen-bonded to other polar amino acids or to the polypeptide backbone.

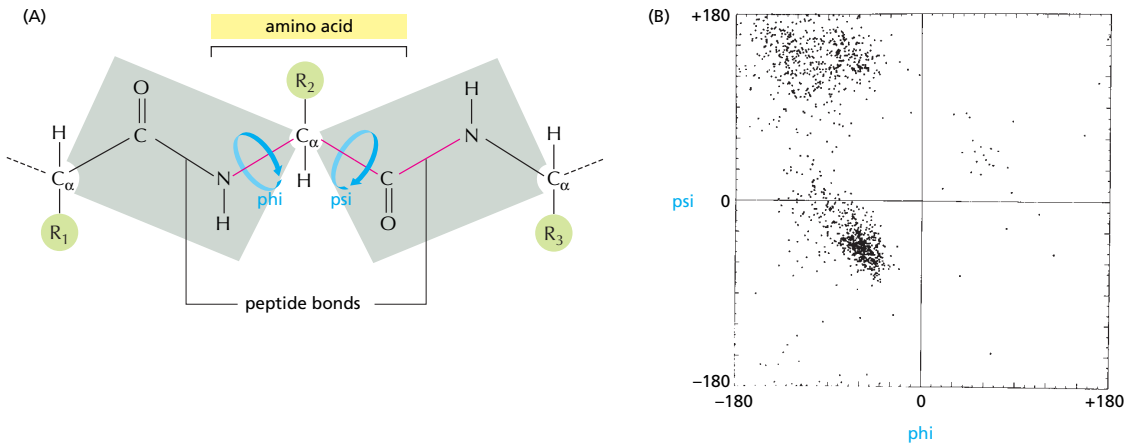
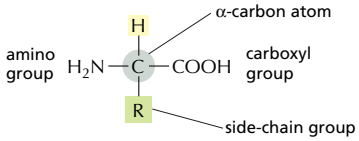



**Figure 3–3 Steric limitations on the bond angles in a polypeptide chain.** (A) Each amino acid contributes three bonds *(red)* to the backbone of the chain. The peptide bond is planar *(gray shading)* and does not permit rotation. By contrast, rotation can occur about the $C_\alpha$–C bond, whose angle of rotation is called psi ($\psi$), and about the N–$C_\alpha$ bond, whose angle of rotation is called phi ($\phi$). By convention, an R group is often used to denote an amino acid side chain *(green circles)*. (B) The conformation of the main-chain atoms in a protein is determined by one pair of $\phi$ and $\psi$ angles for each amino acid; because of steric collisions between atoms within each amino acid, most pairs of $\phi$ and $\psi$ angles do not occur. In this so-called Ramachandran plot, each dot represents an observed pair of angles in a protein. The cluster of dots in the bottom left quadrant represents all of the amino acids that are located in $\alpha$-helix structures (see Figure 3–7A). (B, from J. Richardson, *Adv. Prot. Chem*. 34:174–175, 1981. With permission from Academic Press.)

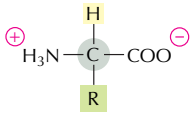
LCY Biotechnology Holding, Inc.
Ex. 1008
Page 9 of 240

## THE AMINO ACID

The general formula of an amino acid is

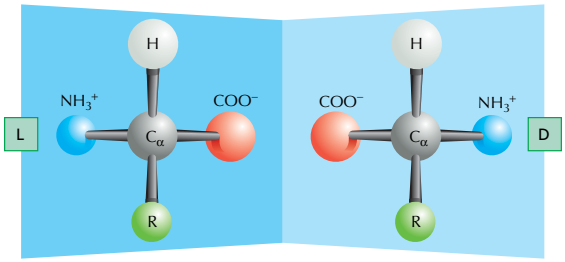



R is commonly one of 20 different side chains. At pH 7 both the amino and carboxyl groups are ionized.




## OPTICAL ISOMERS

The α-carbon atom is asymmetric, which allows for two mirror image (or stereo-) isomers, L and D.




Proteins consist exclusively of L-amino acids.

## FAMILIES OF AMINO ACIDS

The common amino acids are grouped according to whether their side chains are

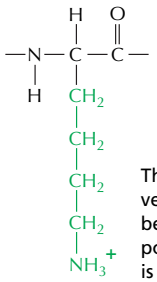
- acidic
- basic
- uncharged polar
- nonpolar

These 20 amino acids are given both three-letter and one-letter abbreviations.

Thus: alanine = Ala = A

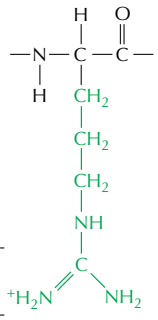
## BASIC SIDE CHAINS

**lysine** (Lys, or K)

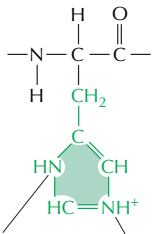


This group is very basic because its positive charge is stabilized by resonance.

**arginine** (Arg, or R)



**histidine** (His, or H)

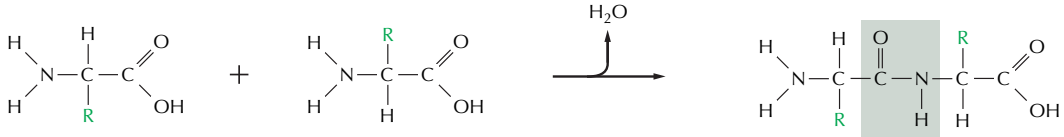


These nitrogens have a relatively weak affinity for an $H^+$ and are only partly positive at neutral pH.

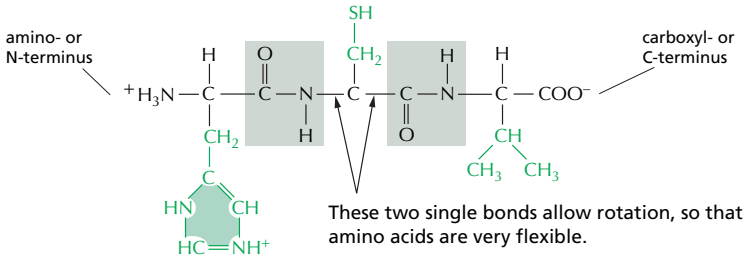
## PEPTIDE BONDS

Amino acids are commonly joined together by an amide linkage, called a peptide bond.

Peptide bond: The four atoms in each *gray box* form a rigid planar unit. There is no rotation around the C–N bond.




Proteins are long polymers of amino acids linked by peptide bonds, and they are always written with the N-terminus toward the left. The sequence of this tripeptide is histidine-cysteine-valine.

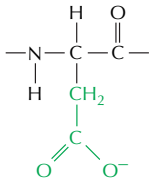



These two single bonds allow rotation, so that long chains of amino acids are very flexible.

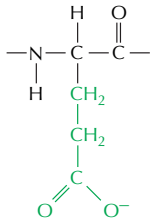
LCY Biotechnology Holding, Inc.
Ex. 1008
Page 10 of 240

## ACIDIC SIDE CHAINS

**aspartic acid**
(Asp, or D)

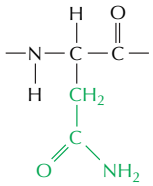


**glutamic acid**
(Glu, or E)

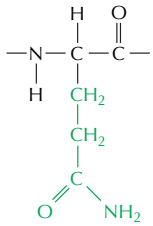


## UNCHARGED POLAR SIDE CHAINS

**asparagine**
(Asn, or N)

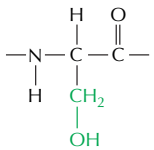


**glutamine**
(Gln, or Q)

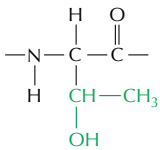


Although the amide N is not charged at neutral pH, it is polar.

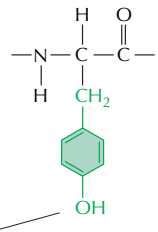
**serine**
(Ser, or S)



**threonine**
(Thr, or T)



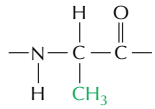
**tyrosine**
(Tyr, or Y)



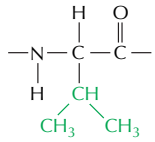
The –OH group is polar.

## NONPOLAR SIDE CHAINS

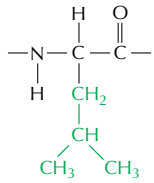
**alanine**
(Ala, or A)



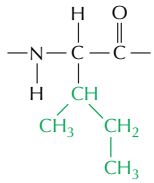
**valine**
(Val, or V)



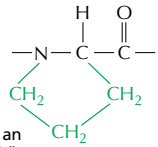
**leucine**
(Leu, or L)



**isoleucine**
(Ile, or I)

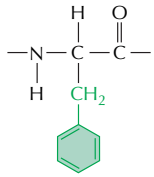


**proline**
(Pro, or P)

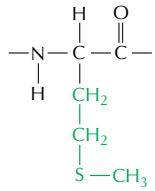


(actually an imino acid)

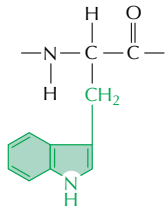
**phenylalanine**
(Phe, or F)



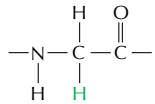
**methionine**
(Met, or M)



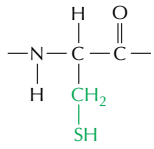
**tryptophan**
(Trp, or W)



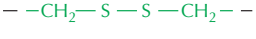
**glycine**
(Gly, or G)



**cysteine**
(Cys, or C)



Disulfide bonds can form between two cysteine side chains in proteins.



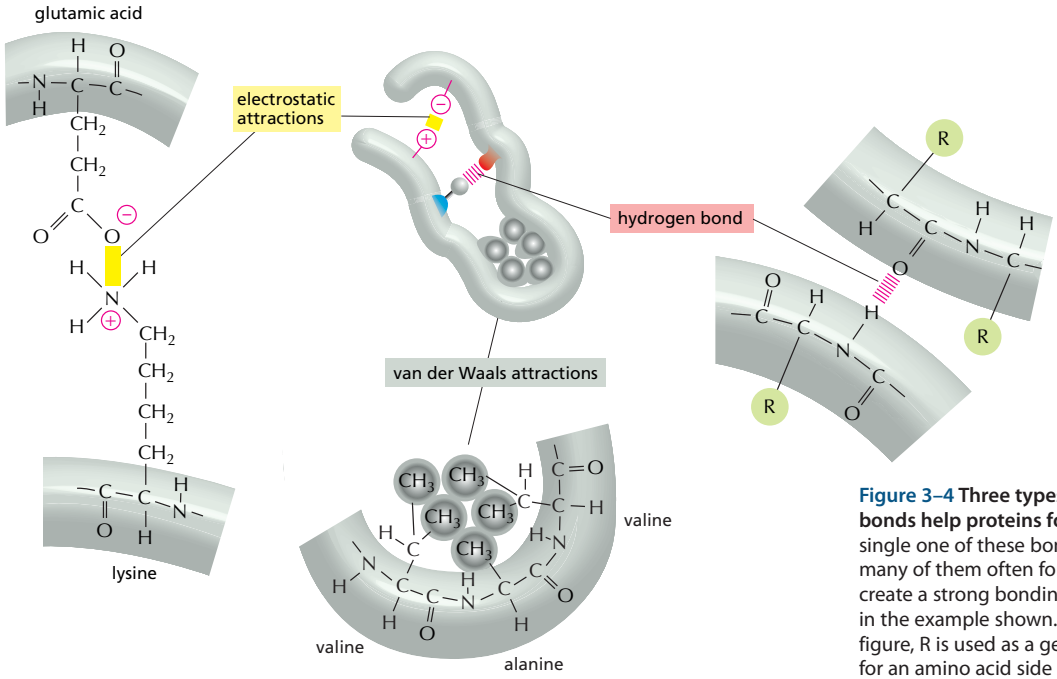



**Figure 3–4 Three types of noncovalent bonds help proteins fold.** Although a single one of these bonds is quite weak, many of them often form together to create a strong bonding arrangement, as in the example shown. As in the previous figure, R is used as a general designation for an amino acid side chain.

## Proteins Fold into a Conformation of Lowest Energy

As a result of all of these interactions, most proteins have a particular three-dimensional structure, which is determined by the order of the amino acids in its chain. The final folded structure, or **conformation**, of any polypeptide chain is generally the one that minimizes its free energy. Biologists have studied protein folding in a test tube by using highly purified proteins. Treatment with certain solvents, which disrupt the noncovalent interactions holding the folded chain together, unfolds, or *denatures*, a protein. This treatment converts the protein into a flexible polypeptide chain that has lost its natural shape. When the denaturing solvent is removed, the protein often refolds spontaneously, or *renatures*, into its original conformation (**Figure 3–6**). This indicates that the amino acid sequence contains all the information needed for specifying the three-dimensional shape of a protein, which is a critical point for understanding cell function.

Each protein normally folds up into a single stable conformation. However, the conformation changes slightly when the protein interacts with other molecules in the cell. This change in shape is often crucial to the function of the protein, as we see later.

Although a protein chain can fold into its correct conformation without outside help, in a living cell special proteins called *molecular chaperones* often assist in protein folding. Molecular chaperones bind to partly folded polypeptide chains and help them progress along the most energetically favorable folding

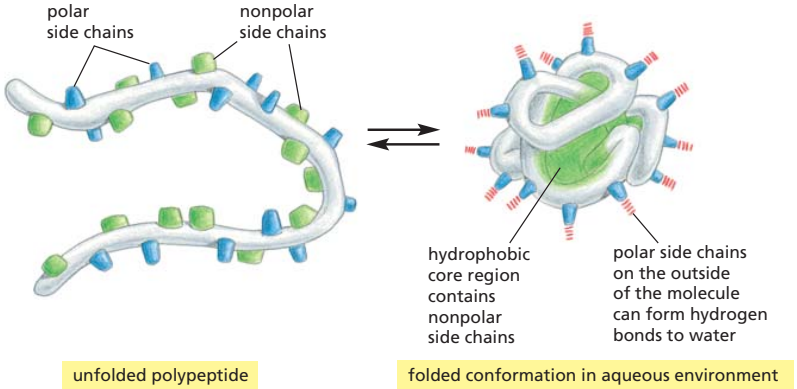



**Figure 3–5 How a protein folds into a compact conformation.** The polar amino acid side chains tend to gather on the outside of the protein, where they can interact with water; the nonpolar amino acid side chains are buried on the inside to form a tightly packed hydrophobic core of atoms that are hidden from water. In this schematic drawing, the protein contains only about 30 amino acids.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 12 of 240

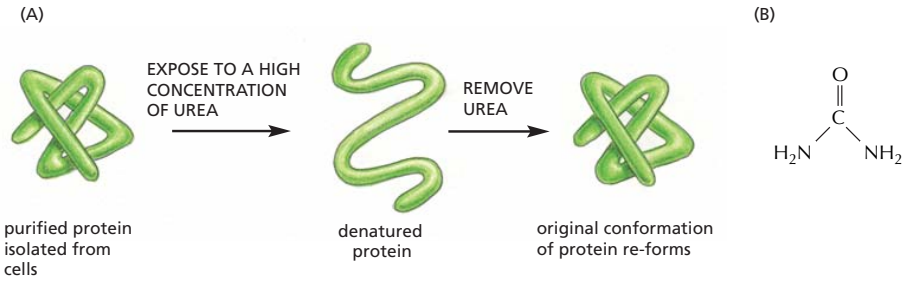



**Figure 3–6 The refolding of a denatured protein.** (A) This type of experiment, first performed more than 40 years ago, demonstrates that a protein's conformation is determined solely by its amino acid sequence. (B) The structure of urea. Urea is very soluble in water and unfolds proteins at high concentrations, where there is about one urea molecule for every six water molecules.

pathway. In the crowded conditions of the cytoplasm, chaperones prevent the temporarily exposed hydrophobic regions in newly synthesized protein chains from associating with each other to form protein aggregates (see p. 388). However, the final three-dimensional shape of the protein is still specified by its amino acid sequence: chaperones simply make the folding process more reliable.

Proteins come in a wide variety of shapes, and they are generally between 50 and 2000 amino acids long. Large proteins usually consist of several distinct *protein domains*—structural units that fold more or less independently of each other, as we discuss below. Since the detailed structure of any protein is complicated, several different representations are used to depict the protein's structure, each emphasizing different features.

**Panel 3–2** (pp. 132–133) presents four different representations of a protein domain called SH2, which has important functions in eucaryotic cells. Constructed from a string of 100 amino acids, the structure is displayed as (A) a polypeptide backbone model, (B) a ribbon model, (C) a wire model that includes the amino acid side chains, and (D) a space-filling model. Each of the three horizontal rows shows the protein in a different orientation, and the image is colored in a way that allows the polypeptide chain to be followed from its N-terminus *(purple)* to its C-terminus *(red)*. <GTGA>

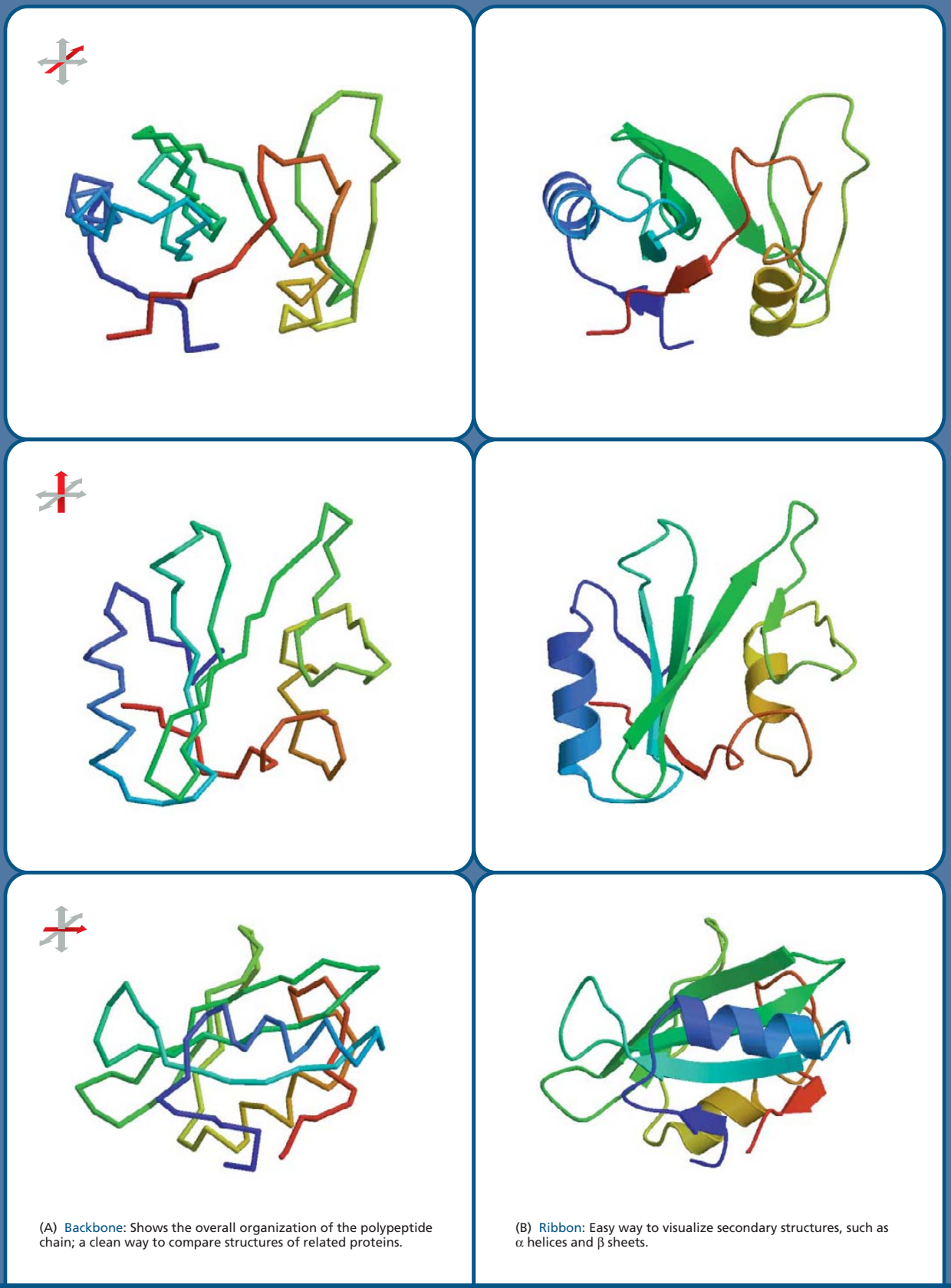
Panel 3–2 shows that a protein's conformation is amazingly complex, even for a structure as small as the SH2 domain. But the description of protein structures can be simplified because they are built up from combinations of several common structural motifs, as we discuss next.

## The α Helix and the β Sheet Are Common Folding Patterns

When we compare the three-dimensional structures of many different protein molecules, it becomes clear that, although the overall conformation of each protein is unique, two regular folding patterns are often found in parts of them. Both patterns were discovered more than 50 years ago from studies of hair and silk. The first folding pattern to be discovered, called the **α helix**, was found in the protein *α-keratin*, which is abundant in skin and its derivatives—such as hair, nails, and horns. Within a year of the discovery of the α helix, a second folded structure, called a **β sheet**, was found in the protein *fibroin*, the major constituent of silk. These two patterns are particularly common because they result from hydrogen-bonding between the N–H and C=O groups in the polypeptide backbone, without involving the side chains of the amino acids. Thus, many different amino acid sequences can form them. In each case, the protein chain adopts a regular, repeating conformation. **Figure 3–7** shows these two conformations, as well as the abbreviations that are used to denote them in ribbon models of proteins.

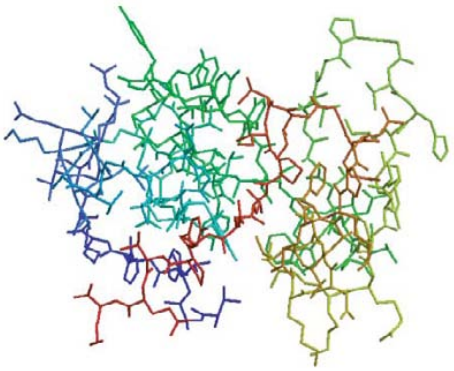
The core of many proteins contains extensive regions of β sheet. As shown in **Figure 3–8**, these β sheets can form either from neighboring polypeptide chains that run in the same orientation (parallel chains) or from a polypeptide chain that folds back and forth upon itself, with each section of the chain running in the direction opposite to that of its immediate neighbors (antiparallel chains). Both types of β sheet produce a very rigid structure, held together by hydrogen bonds that connect the peptide bonds in neighboring chains (see Figure 3–7D).

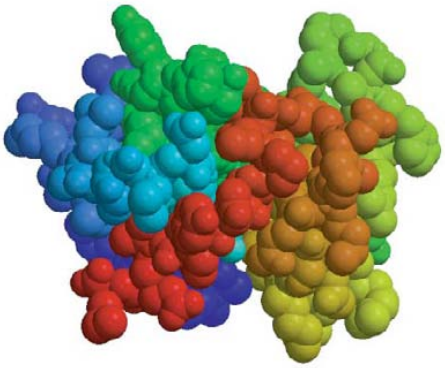
LCY Biotechnology Holding, Inc.
Ex. 1008
Page 13 of 240

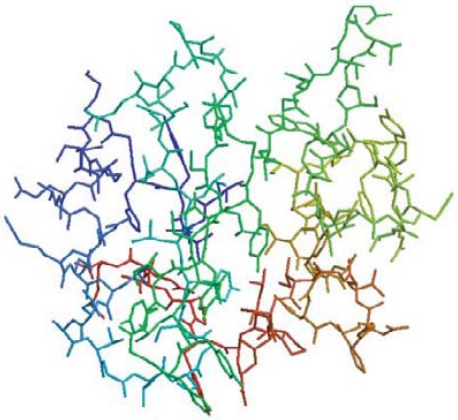


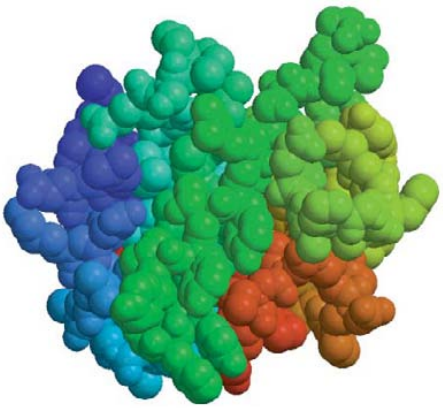
(A) Backbone: Shows the overall organization of the polypeptide chain; a clean way to compare structures of related proteins.

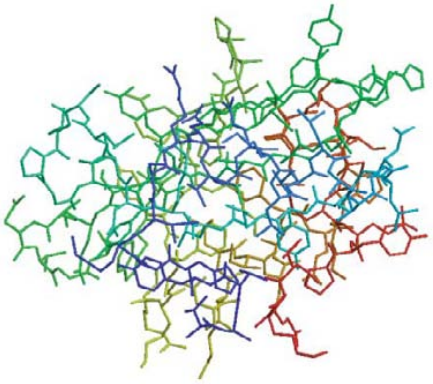
(B) Ribbon: Easy way to visualize secondary structures, such as α helices and β sheets.



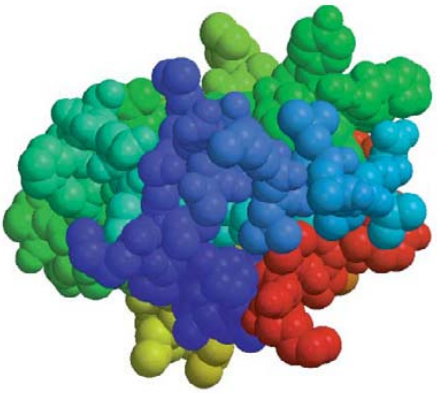








(C) Wire: Highlights side chains and their relative proximities; useful for predicting which amino acids might be involved in a protein's activity, particularly if the protein is an enzyme.



(Courtesy of David Lawson.)

(D) Space-filling: Provides contour map of the protein; gives a feel for the shape of the protein and shows which amino acid side chains are exposed on its surface. Shows how the protein might look to a small molecule, such as water, or to another protein.

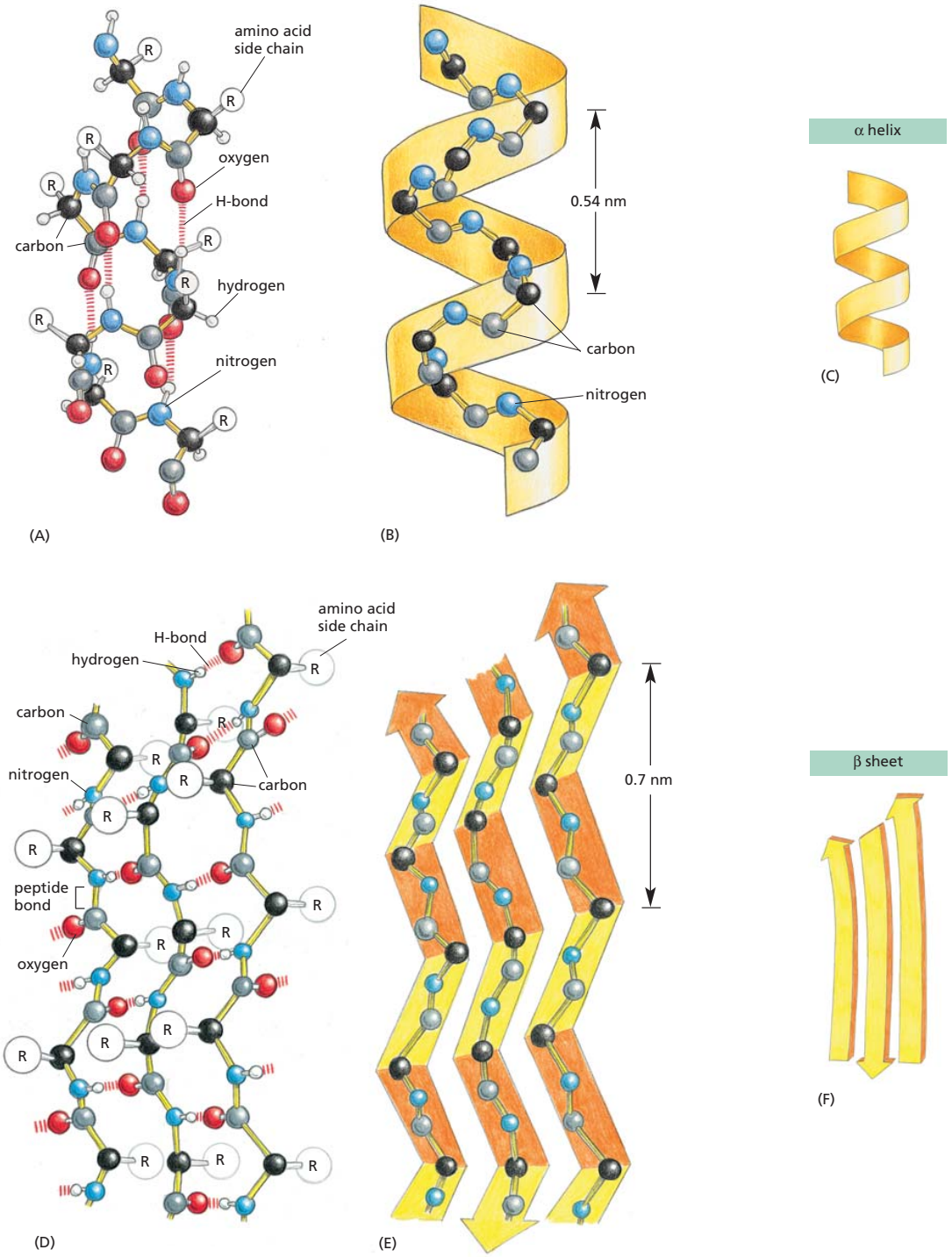



**Figure 3–7 The regular conformation of the polypeptide backbone in the α helix and the β sheet.** <GTAG> <TGCT> (A, B, and C) The α helix. The N–H of every peptide bond is hydrogen-bonded to the C=O of a neighboring peptide bond located four peptide bonds away in the same chain. Note that all of the N–H groups point up in this diagram and that all of the C=O groups point down (toward the C-terminus); this gives a polarity to the helix, with the C-terminus having a partial negative and the N-terminus a partial positive charge. (D, E, and F) The β sheet. In this example, adjacent peptide chains run in opposite (antiparallel) directions. Hydrogen-bonding between peptide bonds in different strands holds the individual polypeptide chains (strands) together in a β sheet, and the amino acid side chains in each strand alternately project above and below the plane of the sheet. (A) and (D) show all the atoms in the polypeptide backbone, but the amino acid side chains are truncated and denoted by R. In contrast, (B) and (E) show the backbone atoms only, while (C) and (F) display the shorthand symbols that are used to represent the α helix and the β sheet in ribbon drawings of proteins (see Panel 3–2B).

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 16 of 240

An α helix is generated when a single polypeptide chain twists around on itself to form a rigid cylinder. A hydrogen bond forms between every fourth peptide bond, linking the C=O of one peptide bond to the N–H of another (see Figure 3–7A). This gives rise to a regular helix with a complete turn every 3.6 amino acids. Note that the protein domain illustrated in Panel 3–2 contains two α helices, as well as a three-stranded antiparallel β sheet.

Regions of α helix are especially abundant in proteins located in cell membranes, such as transport proteins and receptors. As we discuss in Chapter 10, those portions of a transmembrane protein that cross the lipid bilayer usually cross as an α helix composed largely of amino acids with nonpolar side chains. The polypeptide backbone, which is hydrophilic, is hydrogen-bonded to itself in the α helix and shielded from the hydrophobic lipid environment of the membrane by its protruding nonpolar side chains (see also Figure 3–78).

In other proteins, α helices wrap around each other to form a particularly stable structure, known as a **coiled-coil**. This structure can form when the two (or in some cases three) α helices have most of their nonpolar (hydrophobic) side chains on one side, so that they can twist around each other with these side chains facing inward (**Figure 3–9**). Long rodlike coiled-coils provide the structural framework for many elongated proteins. Examples are α-keratin, which forms the intracellular fibers that reinforce the outer layer of the skin and its appendages, and the myosin molecules responsible for muscle contraction.

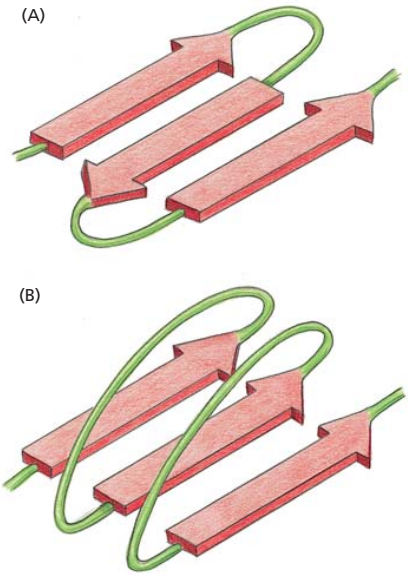



**Figure 3–8 Two types of β sheet structures.** (A) An antiparallel β sheet (see Figure 3–7D). (B) A parallel β sheet. Both of these structures are common in proteins.

## Protein Domains Are Modular Units from which Larger Proteins Are Built

Even a small protein molecule is built from thousands of atoms linked together by precisely oriented covalent and noncovalent bonds, and it is extremely difficult to visualize such a complicated structure without a three-dimensional display. For

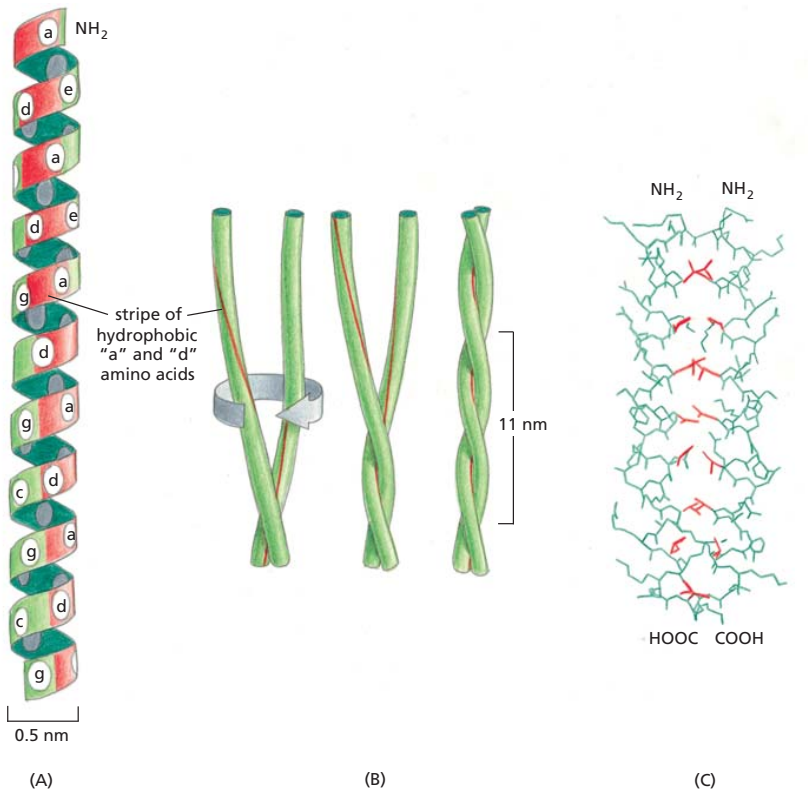



**Figure 3–9 A coiled-coil.** <CGGA> (A) A single α helix, with successive amino acid side chains labeled in a sevenfold sequence, "abcdefg" (*from bottom to top*). Amino acids "a" and "d" in such a sequence lie close together on the cylinder surface, forming a "stripe" (*red*) that winds slowly around the α helix. Proteins that form coiled-coils typically have nonpolar amino acids at positions "a" and "d." Consequently, as shown in (B), the two α helices can wrap around each other with the nonpolar side chains of one α helix interacting with the nonpolar side chains of the other, while the more hydrophilic amino acid side chains are left exposed to the aqueous environment. (C) The atomic structure of a coiled-coil determined by x-ray crystallography. The *red* side chains are nonpolar.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 17 of 240

this reason, biologists use various graphic and computer-based aids. A DVD that accompanies this book contains computer-generated images of selected proteins, displayed and rotated on the screen in a variety of formats.

Biologists distinguish four levels of organization in the structure of a protein. The amino acid sequence is known as the **primary structure**. Stretches of polypeptide chain that form α helices and β sheets constitute the protein's **secondary structure**. The full three-dimensional organization of a polypeptide chain is sometimes referred to as the **tertiary structure**, and if a particular protein molecule is formed as a complex of more than one polypeptide chain, the complete structure is designated as the **quaternary structure**.

Studies of the conformation, function, and evolution of proteins have also revealed the central importance of a unit of organization distinct from these four. This is the **protein domain**, a substructure produced by any part of a polypeptide chain that can fold independently into a compact, stable structure. A domain usually contains between 40 and 350 amino acids, and it is the modular unit from which many larger proteins are constructed.

The different domains of a protein are often associated with different functions. **Figure 3–10** shows an example—the Src protein kinase, which functions in signaling pathways inside vertebrate cells (Src is pronounced "sarc"). This protein is considered to have three domains: the SH2 and SH3 domains have regulatory roles, while the C-terminal domain is responsible for the kinase catalytic activity. Later in the chapter, we shall return to this protein, in order to explain how proteins can form molecular switches that transmit information throughout cells.

**Figure 3–11** presents ribbon models of three differently organized protein domains. As these examples illustrate, the polypeptide chain tends to cross the entire domain before making a sharp turn at the surface. The central core of a domain can be constructed from α helices, from β sheets, or from various combinations of these two fundamental folding elements. <CAGT>

The smallest protein molecules contain only a single domain, whereas larger proteins can contain as many as several dozen domains, often connected to each other by short, relatively unstructured lengths of polypeptide chain.

## Few of the Many Possible Polypeptide Chains Will Be Useful to Cells

Since each of the 20 amino acids is chemically distinct and each can, in principle, occur at any position in a protein chain, there are $20 \times 20 \times 20 \times 20 = 160{,}000$ different possible polypeptide chains four amino acids long, or $20^n$ different possible polypeptide chains $n$ amino acids long. For a typical protein length of

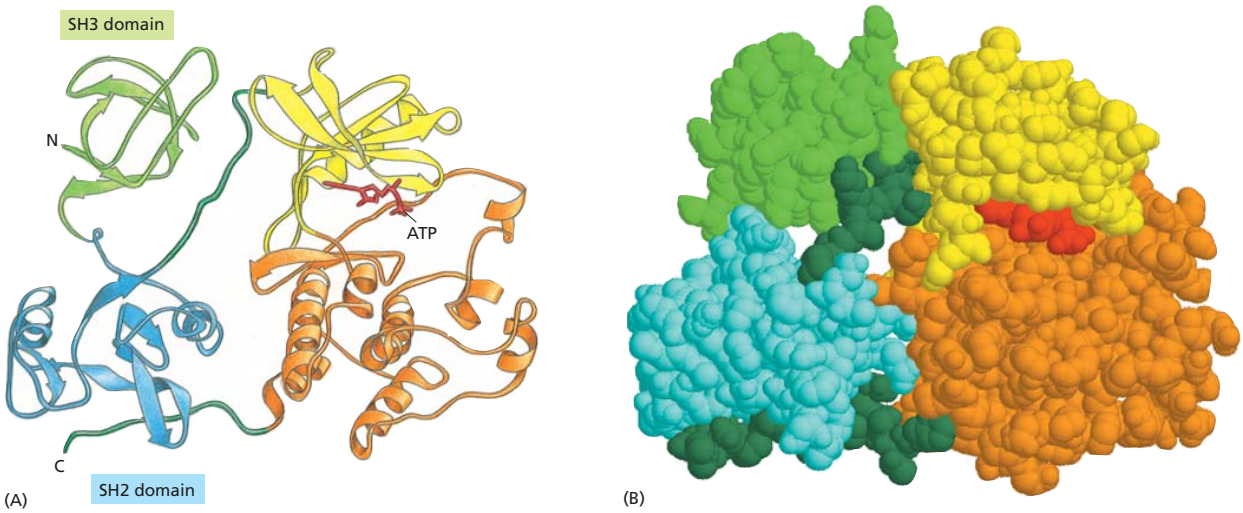



**Figure 3–10 A protein formed from multiple domains.** In the Src protein shown, a C-terminal domain with two lobes (*yellow* and *orange*) forms a protein kinase enzyme, while the SH2 and SH3 domains perform regulatory functions. (A) A ribbon model, with ATP substrate in *red*. (B) A spacing-filling model, with ATP substrate in *red*. Note that the site that binds ATP is positioned at the interface of the two lobes that form the kinase. The detailed structure of the SH2 domain is illustrated in Panel 3–2 (pp. 132–133).

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 18 of 240

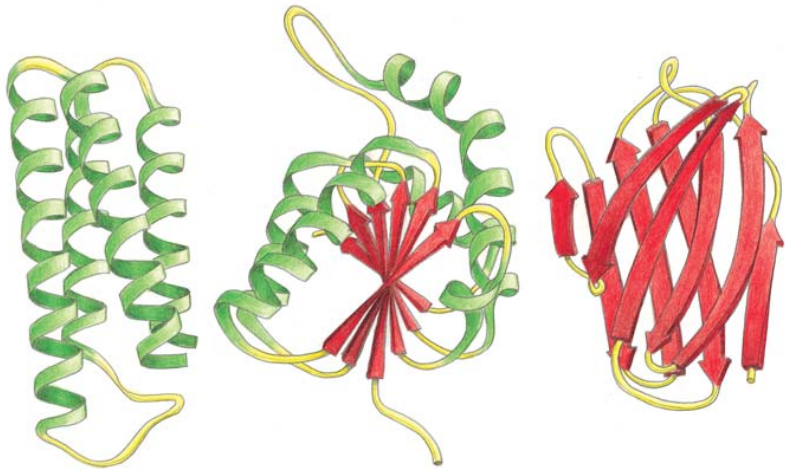



**Figure 3–11 Ribbon models of three different protein domains.** (A) Cytochrome $b_{562}$, a single-domain protein involved in electron transport in mitochondria. This protein is composed almost entirely of α helices. (B) The NAD-binding domain of the enzyme lactic dehydrogenase, which is composed of a mixture of α helices and parallel β sheets. (C) The variable domain of an immunoglobulin (antibody) light chain, composed of a sandwich of two antiparallel β sheets. In these examples, the α helices are shown in *green*, while strands organized as β sheets are denoted by *red arrows*.

Note how the polypeptide chain generally traverses back and forth across the entire domain, making sharp turns only at the protein surface. It is the protruding loop regions *(yellow)* that often form the binding sites for other molecules. (Adapted from drawings courtesy of Jane Richardson.)

about 300 amino acids, a cell could theoretically make more than $10^{390}$ ($20^{300}$) different polypeptide chains. This is such an enormous number that to produce just one molecule of each kind would require many more atoms than exist in the universe.

Only a very small fraction of this vast set of conceivable polypeptide chains would adopt a single, stable three-dimensional conformation—by some estimates, less than one in a billion. And yet the vast majority of proteins present in cells adopt unique and stable conformations. How is this possible? The answer lies in natural selection. A protein with an unpredictably variable structure and biochemical activity is unlikely to help the survival of a cell that contains it. Such proteins would therefore have been eliminated by natural selection through the enormously long trial-and-error process that underlies biological evolution.

Because evolution has selected for protein function in living organisms, the amino acid sequence of most present-day proteins is such that a single conformation is extremely stable. In addition, this conformation has its chemical properties finely tuned to enable the protein to perform a particular catalytic or structural function in the cell. Proteins are so precisely built that the change of even a few atoms in one amino acid can sometimes disrupt the structure of the whole molecule so severely that all function is lost.

## Proteins Can Be Classified into Many Families

Once a protein had evolved that folded up into a stable conformation with useful properties, its structure could be modified during evolution to enable it to perform new functions. This process has been greatly accelerated by genetic mechanisms that occasionally duplicate genes, allowing one gene copy to evolve independently to perform a new function (discussed in Chapter 4). This type of event has occurred quite often in the past; as a result, many present-day proteins can be grouped into protein families, each family member having an amino acid sequence and a three-dimensional conformation that resemble those of the other family members.

Consider, for example, the *serine proteases*, a large family of protein-cleaving (proteolytic) enzymes that includes the digestive enzymes chymotrypsin, trypsin, and elastase, and several proteases involved in blood clotting. When the protease portions of any two of these enzymes are compared, parts of their amino acid sequences are found to match. The similarity of their three-dimensional conformations is even more striking: most of the detailed twists and turns in their polypeptide chains, which are several hundred amino acids long, are virtually identical (**Figure 3–12**). The many different serine proteases nevertheless

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 19 of 240

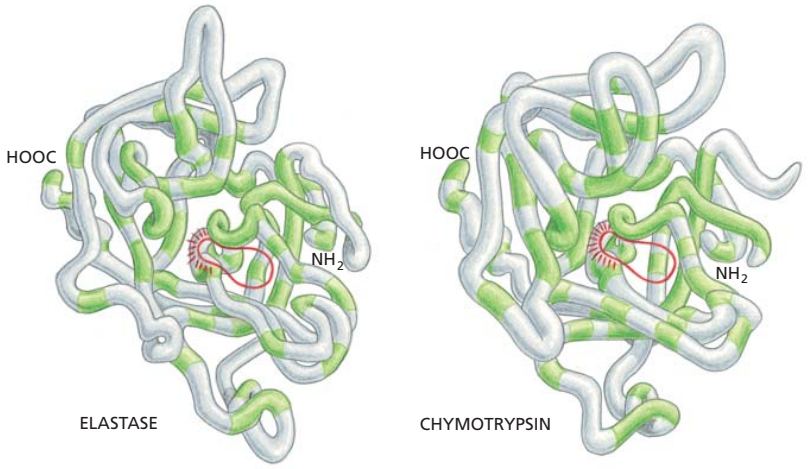


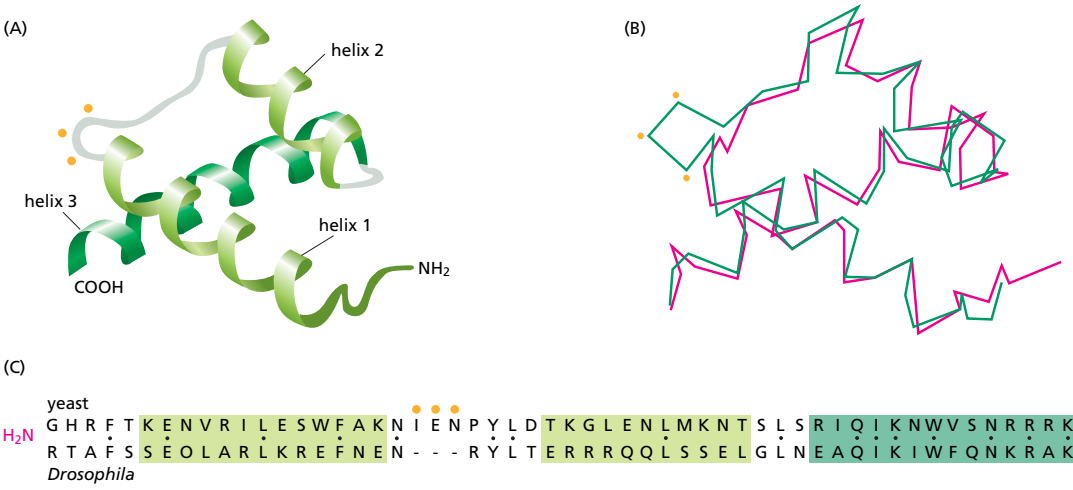

**Figure 3–12 A comparison of the conformations of two serine proteases.** The backbone conformations of elastase and chymotrypsin. Although only those amino acids in the polypeptide chain shaded in *green* are the same in the two proteins, the two conformations are very similar nearly everywhere. The active site of each enzyme is circled in *red;* this is where the peptide bonds of the proteins that serve as substrates are bound and cleaved by hydrolysis. The serine proteases derive their name from the amino acid serine, whose side chain is part of the active site of each enzyme and directly participates in the cleavage reaction.

have distinct enzymatic activities, each cleaving different proteins or the peptide bonds between different types of amino acids. Each therefore performs a distinct function in an organism.

The story we have told for the serine proteases could be repeated for hundreds of other protein families. In general, the structure of the different members of a protein family has been more highly conserved than has the amino acid sequence. In many cases, the amino acid sequences have diverged so far that we cannot be certain of a family relationship between two proteins without determining their three-dimensional structures. The yeast α2 protein and the *Drosophila* engrailed protein, for example, are both gene regulatory proteins in the homeodomain family (discussed in Chapter 7). Because they are identical in only 17 of their 60 amino acid residues, their relationship became certain only by comparing their three-dimensional structures (**Figure 3–13**). Many similar examples show that two proteins with more than 25% identity in their amino acid sequences usually share the same overall structure.

The various members of a large protein family often have distinct functions. Some of the amino acid changes that make family members different were no doubt selected in the course of evolution because they resulted in useful changes in biological activity, giving the individual family members the different functional properties they have today. But many other amino acid changes are effectively "neutral," having neither a beneficial nor a damaging effect on the basic structure and function of the protein. In addition, since mutation is a random process, there must also have been many deleterious changes that altered the three-dimensional structure of these proteins sufficiently to harm them.

**Figure 3–13 A comparison of a class of DNA-binding domains, called homeodomains, in a pair of proteins from two organisms separated by more than a billion years of evolution.** (A) A ribbon model of the structure common to both proteins. (B) A trace of the α-carbon positions. The three-dimensional structures shown were determined by x-ray crystallography for the yeast α2 protein *(green)* and the *Drosophila* engrailed protein *(red)*. (C) A comparison of amino acid sequences for the region of the proteins shown in (A) and (B). *Black dots* mark sites with identical amino acids. *Orange dots* indicate the position of a three amino acid insert in the α2 protein. (Adapted from C. Wolberger et al., *Cell* 67:517–528, 1991. With permission from Elsevier.)




LCY Biotechnology Holding, Inc.
Ex. 1008
Page 20 of 240

Such faulty proteins would have been lost whenever the individual organisms making them were at enough of a disadvantage to be eliminated by natural selection.

Protein families are readily recognized when the genome of any organism is sequenced; for example, the determination of the DNA sequence for the entire human genome has revealed that we contain about 24,000 protein-coding genes. Through sequence comparisons, we can assign the products of about 40 percent of these genes to known protein structures, belonging to more than 500 different protein families. Most of the proteins in each family have evolved to perform somewhat different functions, as for the enzymes elastase and chymotrypsin illustrated previously in Figure 3–12. These are sometimes called *paralogs* to distinguish them from the corresponding proteins in different organisms (*orthologs*, such as mouse and human elastase).

As described in Chapter 8, because of the powerful techniques of x-ray crystallography and nuclear magnetic resonance (NMR), we now know the three-dimensional shapes, or conformations, of more than 20,000 proteins. <GGCC> By carefully comparing the conformations of these proteins, structural biologists (that is, experts on the structure of biological molecules) have concluded that there are a limited number of ways in which protein domains fold up in nature—maybe as few as 2000. The structures for about 800 of these protein folds have thus far been determined. These known folds tend to be those most represented in the universe of protein structures: for example, 50 folds account for nearly three-fourths of the domain families with predicted structures. A complete catalog of the most significant protein folds that exist in living organisms would therefore seem to be within our reach.

## Sequence Searches Can Identify Close Relatives

The present database of known protein sequences contains more than ten million entries, and it is growing very rapidly as more and more genomes are sequenced—revealing huge numbers of new genes that encode proteins. Powerful computer search programs are available that allow us to compare each newly discovered protein with this entire database, looking for possible relatives. Many proteins whose genes have evolved from a common ancestral gene can be identified by the discovery of statistically significant similarities in amino acid sequences.

With such a large number of proteins in the database, the search programs find many nonsignificant matches, resulting in a background noise level that makes it very difficult to pick out all but the closest relatives. Generally speaking, one requires a 30% identity in sequence to consider that two proteins match. However, we know the function of many short signature sequences ("fingerprints"), and these are widely used to find more distant relationships (**Figure 3–14**).

Protein comparisons are important because related structures often imply related functions. Many years of experimentation can be saved by discovering that a new protein has an amino acid sequence similarity with a protein of known function. Such sequence relationships, for example, first indicated that certain genes that cause mammalian cells to become cancerous are protein kinases. In the same way, many of the proteins that control pattern formation during the embryonic development of the fruit fly *Drosophila* were quickly recognized to be gene regulatory proteins.

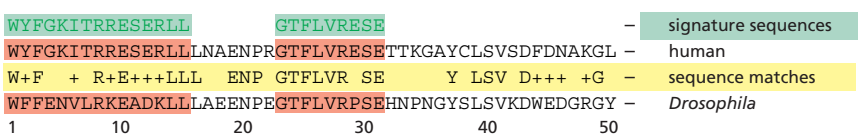



**Figure 3–14 The use of short signature sequences to find related protein domains.** The two short sequences of 15 and 9 amino acids shown *(green)* can be used to search large databases for a protein domain that is found in many proteins, the SH2 domain. Here, the first 50 amino acids of the SH2 domain of 100 amino acids is compared for the human and *Drosophila* Src protein (see Figure 3–10). In the computer-generated sequence comparison *(yellow row),* exact matches between the human and *Drosophila* proteins are noted by the one-letter abbreviation for the amino acid; the positions with a similar but nonidentical amino acid are denoted by +, and nonmatches are blank. In this diagram, wherever one or both proteins contain an exact match to a position in the *green* sequences, both aligned sequences are colored *red*.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 21 of 240

## Some Protein Domains Form Parts of Many Different Proteins

As previously stated, most proteins are composed of a series of protein domains, in which different regions of the polypeptide chain have folded independently to form compact structures. Such multidomain proteins are believed to have originated from the accidental joining of the DNA sequences that encode each domain, creating a new gene. Novel binding surfaces have often been created at the juxtaposition of domains, and many of the functional sites where proteins bind to small molecules are found to be located there. In an evolutionary process called *domain shuffling*, many large proteins have evolved through the joining of preexisting domains in new combinations (**Figure 3–15**).

A subset of protein domains have been especially mobile during evolution; these seem to have particularly versatile structures and are sometimes referred to as *protein modules*. The structure of one such module, the SH2 domain, was illustrated in Panel 3–2 (pp. 132–133). Some other abundant protein domains are illustrated in **Figure 3–16**.

Each of the domains shown has a stable core structure formed from strands of β sheet, from which less-ordered loops of polypeptide chain protrude *(green)*. The loops are ideally situated to form binding sites for other molecules, as most clearly demonstrated for the immunoglobulin fold, which forms the basis for antibody molecules (see Figure 3–41). Most likely, such β-sheet-based domains have achieved their evolutionary success because they provide a convenient framework for the generation of new binding sites for ligands through small changes to their protruding loops.

A second feature of these protein domains that explains their utility is the ease with which they can be integrated into other proteins. Five of the six domains illustrated in Figure 3–16 have their N- and C-terminal ends at opposite poles of the domain. When the DNA encoding such a domain undergoes tandem duplication, which is not unusual in the evolution of genomes (discussed in Chapter 4), the duplicated modules with this "in-line" arrangement

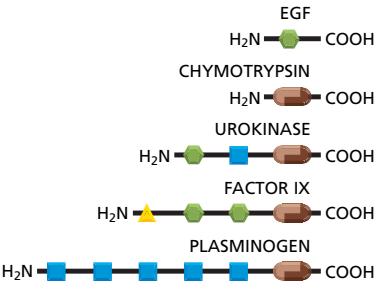



**Figure 3–15 Domain shuffling.** An extensive shuffling of blocks of protein sequence (protein domains) has occurred during protein evolution. Those portions of a protein denoted by the same shape and color in this diagram are evolutionarily related. Serine proteases like chymotrypsin are formed from two domains *(brown)*. In the three other proteases shown, which are highly regulated and more specialized, these two protease domains are connected to one or more domains that are similar to domains found in epidermal growth factor (EGF; *green),* to a calcium-binding protein *(yellow),* or to a "kringle" domain *(blue)* that contains three internal disulfide bridges. Chymotrypsin is illustrated in Figure 3–12.

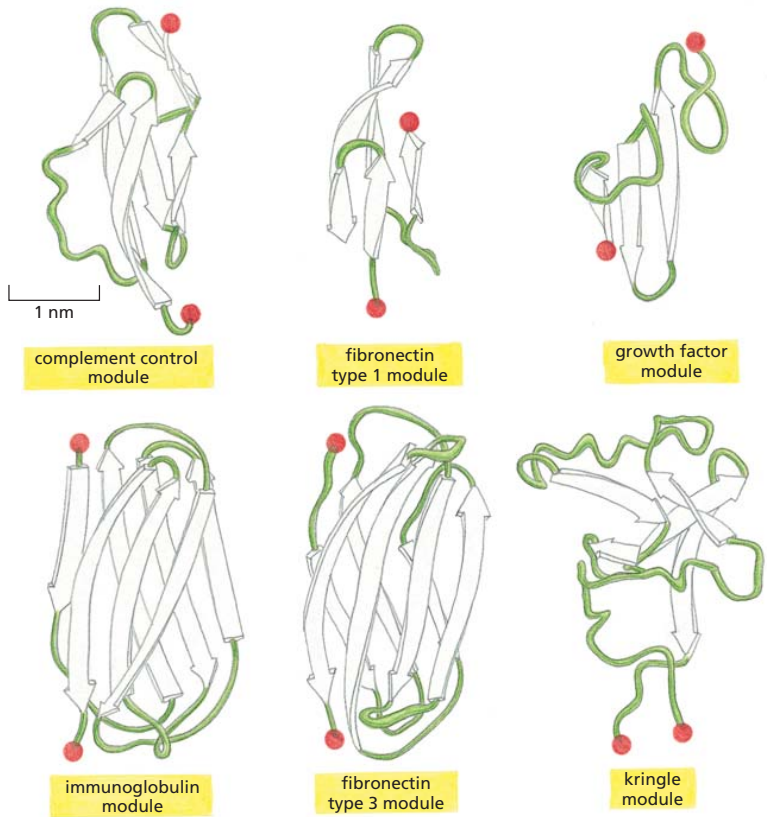



**Figure 3–16 The three-dimensional structures of some protein modules.** In these ribbon diagrams, β-sheet strands are shown as *arrows,* and the N- and C-termini are indicated by *red spheres.* (Adapted from M. Baron, D.G. Norman and I.D. Campbell, *Trends Biochem. Sci.* 16:13–17, 1991, with permission from Elsevier, and D.J. Leahy et al., *Science* 258:987–991, 1992, with permission from AAAS.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 22 of 240

can be readily linked in series to form extended structures—either with themselves or with other in-line domains (**Figure 3–17**). Stiff extended structures composed of a series of domains are especially common in extracellular matrix molecules and in the extracellular portions of cell-surface receptor proteins. Other modules, including the SH2 domain and the kringle domain illustrated in Figure 3–16, are of a "plug-in" type, with their N- and C-termini close together. After genomic rearrangements, such modules are usually accommodated as an insertion into a loop region of a second protein.

A comparison of the relative frequency of domain utilization in different eucaryotes reveals that, for many common domains, such as protein kinases, this frequency is similar in organisms as diverse as yeast, plants, worms, flies, and humans (**Figure 3–18**). But there are some notable exceptions, such as the Major Histocompatibility Complex (MHC) antigen-recognition domain (see Figure 25–52) that is present in 57 copies in humans, but absent in the other four organisms just mentioned. Such domains presumably have specialized functions that are not shared with the other eucaryotes, being strongly selected for during evolution so as to give rise to the multiple copies observed. Similarly, a domain like SH2 that shows an unusual increase in its numbers in higher eucaryotes might be assumed to be especially useful for multicellularity (compare the multicellular organisms with yeast in Figure 3–18).

## Certain Pairs of Domains Are Found Together in Many Proteins

We can construct a large table displaying domain usage for each organism whose genome sequence is known. For example, the human genome is estimated to contain about 1000 immunoglobulin domains, 500 protein kinase domains, 250 DNA-binding homeodomains, 300 SH3 domains, and 120 SH2 domains. Important additional information can be derived by comparing the frequencies and arrangements of domains in the more than 100 eucaryotic, bacterial, and archaeal genomes that have been completely sequenced. For example, we find that more than two-thirds of proteins consist of two or more domains, and that the same pairs of domains occur repeatedly in the same relative arrangement in a protein. Although half of all domain families are common to archaea, bacteria, and eucaryotes, only about 5 percent of the two-domain combinations are similarly shared. This pattern suggests that most proteins containing especially useful two-domain combinations arose relatively late in evolution.

The 200 most abundant two-domain combinations occur in about one-fourth of all of the proteins with recognizable domains in the complete data set. It would therefore be very useful to determine the precise three-dimensional structure for at least one protein from each common two-domain combination, so as to reveal how the domains interact in that type of protein structure.

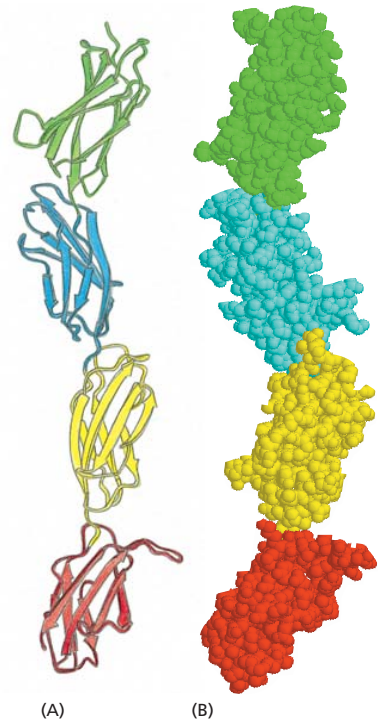



**Figure 3–17 An extended structure formed from a series of in-line protein modules.** Four fibronectin type 3 modules (see Figure 3–16) from the extracellular matrix molecule fibronectin are illustrated in (A) ribbon and (B) space-filling models. (Adapted from D.J. Leahy, I. Aukhil and H.P. Erickson, *Cell* 84:155–164, 1996. With permission from Elsevier.)

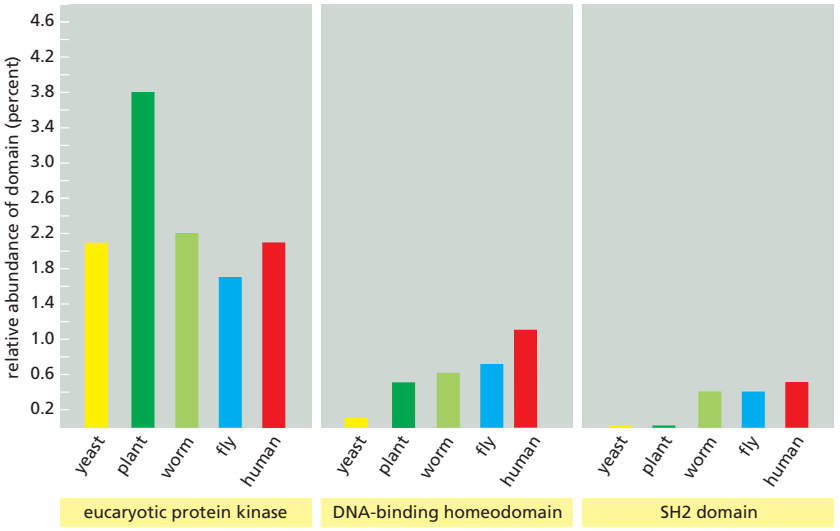



**Figure 3–18 The relative frequencies of three protein domains in five eucaryotic organisms.** The approximate percentages given have been determined by dividing the number of copies of each domain by the total number of distinct proteins thought to be encoded by each organism, as determined from the sequence of its genome. Thus, for SH2 domains in humans, 120/24,000 = 0.005.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 23 of 240

## The Human Genome Encodes a Complex Set of Proteins, Revealing Much That Remains Unknown

The result of sequencing the human genome has been surprising, because it reveals that our chromosomes contain only about 25,000 genes. Based on gene number alone, we would appear to be no more complex than the tiny mustard weed, *Arabidopsis,* and only about 1.3-fold more complex than a nematode worm. The genome sequences also reveal that vertebrates have inherited nearly all of their protein domains from invertebrates—with only 7 percent of identified human domains being vertebrate-specific.

Each of our proteins is on average more complicated, however (**Figure 3–19**). Domain shuffling during vertebrate evolution has given rise to many novel combinations of protein domains, with the result that there are nearly twice as many combinations of domains found in human proteins as in a worm or a fly. Thus, for example, the trypsinlike serine protease domain is linked to at least 18 other types of protein domains in human proteins, whereas it is found covalently joined to only 5 different domains in the worm. This extra variety in our proteins greatly increases the range of protein–protein interactions possible (see Figure 3–82), but how it contributes to making us human is not known.

The complexity of living organisms is staggering, and it is quite sobering to note that we currently lack even the tiniest hint of what the function might be for more than 10,000 of the proteins that have thus far been identified in the human genome. There are certainly enormous challenges ahead for the next generation of cell biologists, with no shortage of fascinating mysteries to solve.

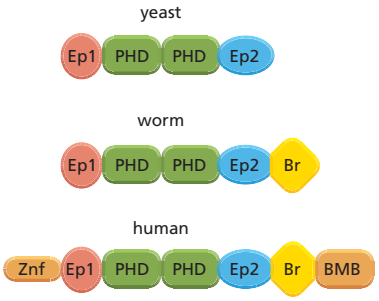



**Figure 3–19 Domain structure of a group of evolutionarily related proteins that are thought to have a similar function.** In general, there is a tendency for the proteins in more complex organisms, such as humans, to contain additional domains—as is the case for the DNA-binding protein compared here.

## Larger Protein Molecules Often Contain More Than One Polypeptide Chain <GCCT>

The same weak noncovalent bonds that enable a protein chain to fold into a specific conformation also allow proteins to bind to each other to produce larger structures in the cell. Any region of a protein's surface that can interact with another molecule through sets of noncovalent bonds is called a **binding site**. A protein can contain binding sites for various large and small molecules. If a binding site recognizes the surface of a second protein, the tight binding of two folded polypeptide chains at this site creates a larger protein molecule with a precisely defined geometry. Each polypeptide chain in such a protein is called a **protein subunit**.

In the simplest case, two identical folded polypeptide chains bind to each other in a "head-to-head" arrangement, forming a symmetric complex of two protein subunits (a *dimer*) held together by interactions between two identical binding sites. The *Cro repressor protein*—a viral gene regulatory protein that binds to DNA to turn viral genes off in an infected bacterial cell—provides an example (**Figure 3–20**). Cells contain many other types of symmetric protein complexes, formed from multiple copies of a single polypeptide chain. The enzyme *neuraminidase*, for example, consists of four identical protein subunits, each bound to the next in a "head-to-tail" arrangement that forms a closed ring (**Figure 3–21**).

Many of the proteins in cells contain two or more types of polypeptide chains. *Hemoglobin*, the protein that carries oxygen in red blood cells, contains

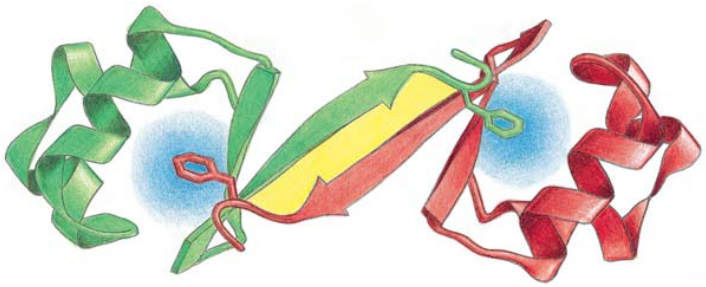


**Figure 3–20 Two identical protein subunits binding together to form a symmetric protein dimer.** The Cro repressor protein from bacteriophage lambda binds to DNA to turn off viral genes. Its two identical subunits bind head-to-head, held together by a combination of hydrophobic forces *(blue)* and a set of hydrogen bonds *(yellow* region). (Adapted from D.H. Ohlendorf, D.E. Tronrud and B.W. Matthews, *J. Mol. Biol*. 280:129–136, 1998. With permission from Academic Press.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 24 of 240

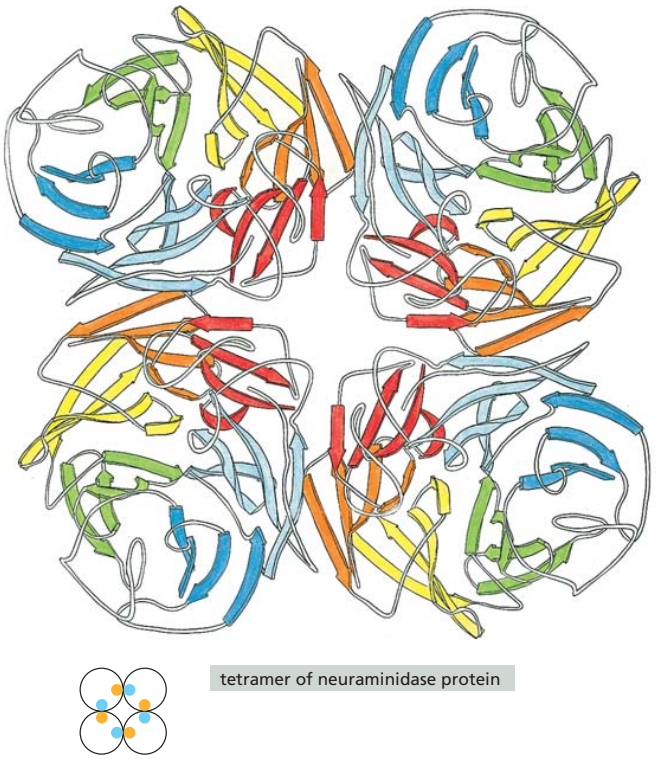



**Figure 3–21 A protein molecule containing multiple copies of a single protein subunit.** The enzyme neuraminidase exists as a ring of four identical polypeptide chains. Each of these chains is formed from six repeats of a four-stranded β sheet, as indicated by the colored arrows. The small diagram shows how the repeated use of the same binding interaction forms the structure.

two identical α-globin subunits and two identical β-globin subunits, symmetrically arranged (**Figure 3–22**). Such multisubunit proteins are very common in cells, and they can be very large. **Figure 3–23** shows a sample of proteins whose exact structures are known, and it compares the sizes and shapes of a few larger proteins with some of the relatively small proteins that we have thus far presented as models.

## Some Proteins Form Long Helical Filaments

Some protein molecules can assemble to form filaments that may span the entire length of a cell. Most simply, a long chain of identical protein molecules can be constructed if each molecule has a binding site complementary to another region of the surface of the same molecule (**Figure 3–24**). An actin filament, for example, is a long helical structure produced from many molecules of the protein *actin* (**Figure 3–25**). Actin is very abundant in eucaryotic cells, where it constitutes one of the major filament systems of the cytoskeleton (discussed in Chapter 16).

Why is a helix such a common structure in biology? As we have seen, biological structures are often formed by linking similar subunits—such as amino acids or protein molecules—into long, repetitive chains. If all the subunits are identical, the neighboring subunits in the chain can often fit together in only one way, adjusting their relative positions to minimize the free energy of the contact between them. As a result, each subunit is positioned in exactly the same way in relation to the next, so that subunit 3 fits onto subunit 2 in the same way that subunit 2 fits onto subunit 1, and so on. Because it is very rare for subunits to join up in a straight line, this arrangement generally results in a helix—a regular structure that resembles a spiral staircase, as illustrated in **Figure 3–26**. Depending on the twist of the staircase, a helix is said to be either right-handed or left-handed (see Figure 3–26E). Handedness is not affected by turning the helix upside down, but it is reversed if the helix is reflected in the mirror.

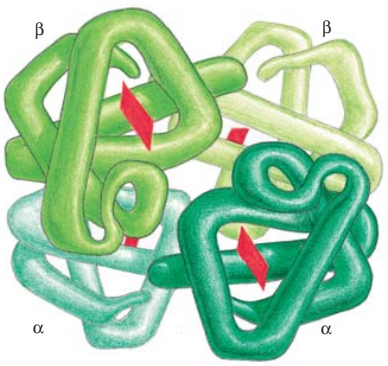



**Figure 3–22 A protein formed as a symmetric assembly of two different subunits.** Hemoglobin is an abundant protein in red blood cells that contains two copies of α-globin and two copies of β-globin. Each of these four polypeptide chains contains a heme molecule *(red)*, which is the site that binds oxygen ($O_2$). Thus, each molecule of hemoglobin in the blood carries four molecules of oxygen.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 25 of 240

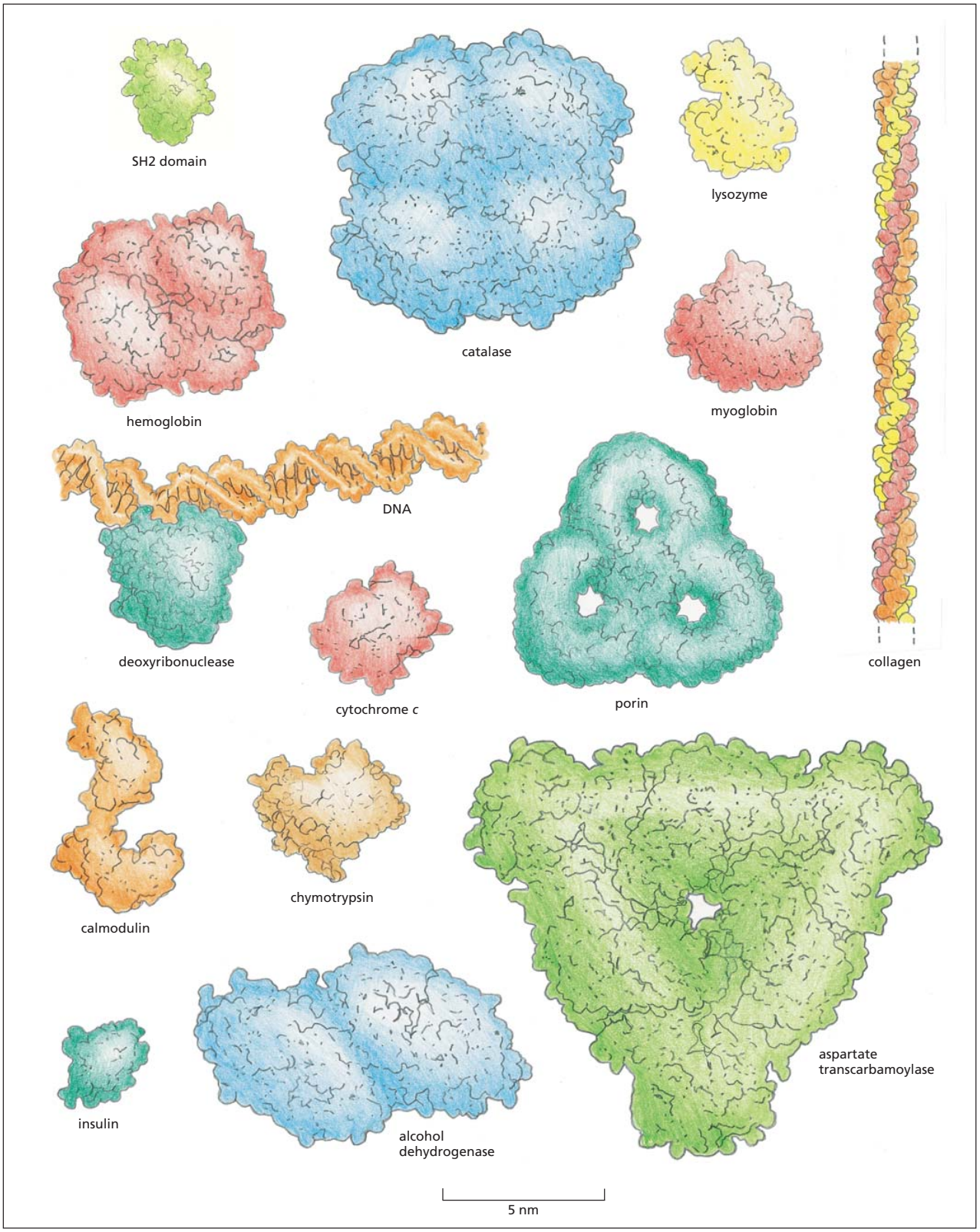



**Figure 3–23 A collection of protein molecules, shown at the same scale.** For comparison, a DNA molecule bound to a protein is also illustrated. These space-filling models represent a range of sizes and shapes. Hemoglobin, catalase, porin, alcohol dehydrogenase, and aspartate transcarbamoylase are formed from multiple copies of subunits. The SH2 domain *(top left)* is presented in detail in Panel 3–2 (pp. 132–133). (Adapted from David S. Goodsell, Our Molecular Nature. New York: Springer-Verlag, 1996. With permission from Springer Science and Business Media.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 26 of 240

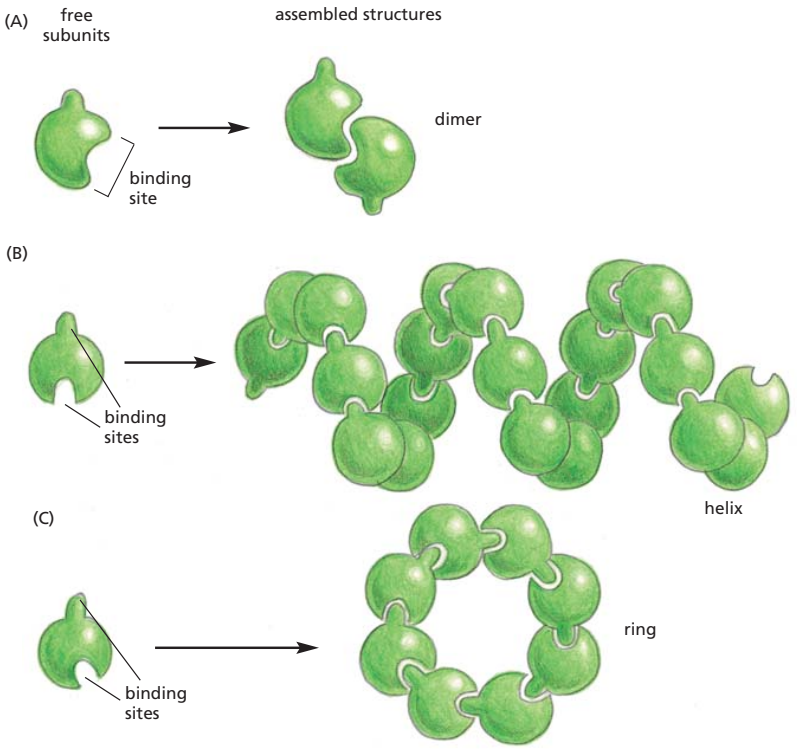



**Figure 3–24 Protein assemblies.** (A) A protein with just one binding site can form a dimer with another identical protein. (B) Identical proteins with two different binding sites often form a long helical filament. (C) If the two binding sites are disposed appropriately in relation to each other, the protein subunits may form a closed ring instead of a helix. (For an example of A, see Figure 3–20; for an example of C, see Figure 3–21.)

Helices occur commonly in biological structures, whether the subunits are small molecules linked together by covalent bonds (for example, the amino acids in an α helix) or large protein molecules that are linked by noncovalent forces (for example, the actin molecules in actin filaments). This is not surprising. A helix is an unexceptional structure, and it is generated simply by placing many similar subunits next to each other, each in the same strictly repeated relationship to the one before—that is, with a fixed rotation followed by a fixed translation along the helix axis, as in a spiral staircase.

## Many Protein Molecules Have Elongated, Fibrous Shapes

Most of the proteins that we have discussed so far are *globular proteins,* in which the polypeptide chain folds up into a compact shape like a ball with an irregular surface. Enzymes tend to be globular proteins: even though many are large and complicated, with multiple subunits, most have an overall rounded shape (see Figure 3–23). In contrast, other proteins have roles in the cell that require each individual protein molecule to span a large distance. These proteins generally have a relatively simple, elongated three-dimensional structure and are commonly referred to as *fibrous proteins*.

One large family of intracellular fibrous proteins consists of α-keratin, introduced when we presented the α helix, and its relatives. Keratin filaments are extremely stable and are the main component in long-lived structures such as hair, horn, and nails. An α-keratin molecule is a dimer of two identical subunits, with the long α helices of each subunit forming a coiled-coil (see Figure 3–9). The coiled-coil regions are capped at each end by globular domains containing binding sites. This enables this class of protein to assemble into ropelike *intermediate filaments*—an important component of the cytoskeleton that creates the cell's internal structural framework (see Figure 16–19).

Fibrous proteins are especially abundant outside the cell, where they are a main component of the gel-like *extracellular matrix* that helps to bind collections of cells together to form tissues. Cells secrete extracellular matrix proteins into their surroundings, where they often assemble into sheets or long fibrils.

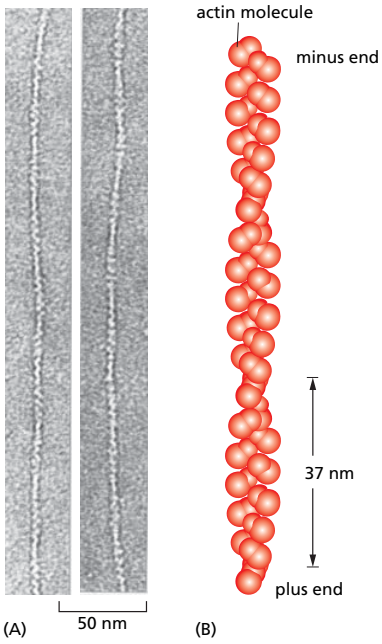
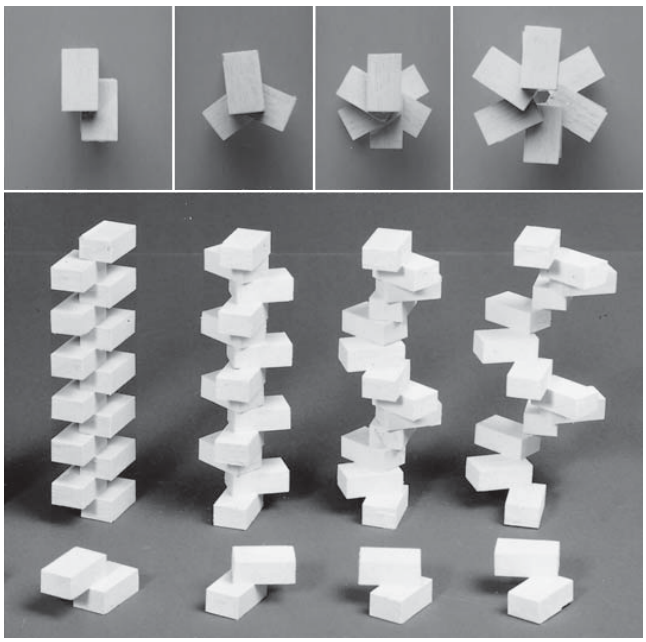



**Figure 3–25 Actin filaments.** (A) Transmission electron micrographs of negatively stained actin filaments. (B) The helical arrangement of actin molecules in an actin filament. (A, courtesy of Roger Craig.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 27 of 240



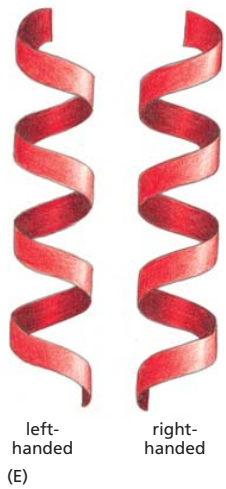
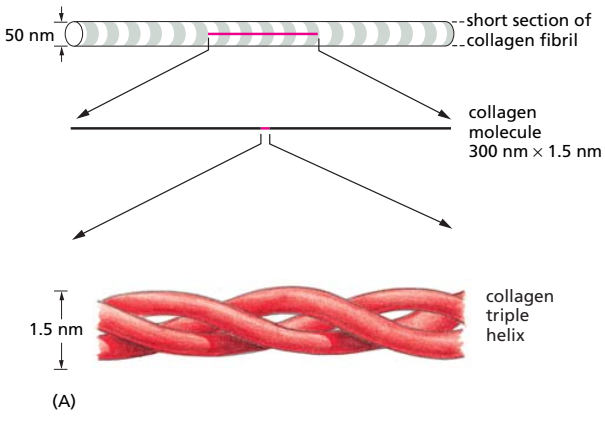




**Figure 3–26 Some properties of a helix.** (A–D) A helix forms when a series of subunits bind to each other in a regular way. At the bottom, the interaction between two subunits is shown; behind them are the helices that result. These helices have two (A), three (B), and six (C and D) subunits per helical turn. The photographs at the top show the arrangement of subunits viewed from directly above the helix. Note that the helix in (D) has a wider path than that in (C), but the same number of subunits per turn. (E) A helix can be either right-handed or left-handed. As a reference, it is useful to remember that standard metal screws, which insert when turned clockwise, are right-handed. Note that a helix retains the same handedness when it is turned upside down.

*Collagen* is the most abundant of these proteins in animal tissues. A collagen molecule consists of three long polypeptide chains, each containing the nonpolar amino acid glycine at every third position. This regular structure allows the chains to wind around one another to generate a long regular triple helix (**Figure 3–27**A). Many collagen molecules then bind to one another side-by-side and end-to-end to create long overlapping arrays—thereby generating the extremely tough collagen fibrils that give connective tissues their tensile strength, as described in Chapter 19.

## Many Proteins Contain a Surprisingly Large Amount of Unstructured Polypeptide Chain

It has been well known for a long time that, in complete contrast to collagen, another abundant protein in the extracellular matrix, *elastin*, is formed as a highly disordered polypeptide. This disorder is essential for elastin's function. Its



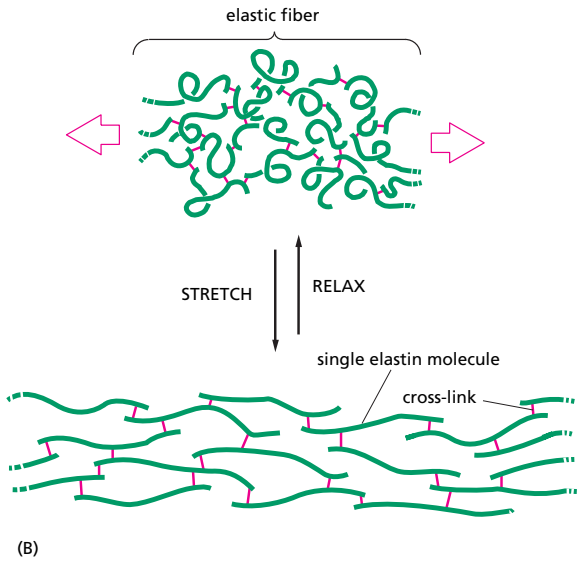




**Figure 3–27 Collagen and elastin.** (A) Collagen is a triple helix formed by three extended protein chains that wrap around one another *(bottom)*. Many rodlike collagen molecules are cross-linked together in the extracellular space to form unextendable collagen fibrils *(top)* that have the tensile strength of steel. The striping on the collagen fibril is caused by the regular repeating arrangement of the collagen molecules within the fibril. (B) Elastin polypeptide chains are cross-linked together to form rubberlike, elastic fibers. Each elastin molecule uncoils into a more extended conformation when the fiber is stretched and recoils spontaneously as soon as the stretching force is relaxed.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 28 of 240

relatively loose and unstructured polypeptide chains are covalently cross-linked to produce a rubberlike elastic meshwork that can be reversibly pulled from one conformation to another, as illustrated in Figure 3–27B. The elastic fibers that result enable skin and other tissues, such as arteries and lungs, to stretch and recoil without tearing.

Intrinsically unstructured regions of proteins are quite frequent in nature, having important functions in the interior of cells. As we have already seen, proteins use the short loops of polypeptide chain that generally protrude from the core region of protein domains to bind other molecules. Similarly, many proteins have much longer regions of unstructured amino acid sequences that interact with another molecule (often DNA or a protein), undergoing a structural transition to a specific folded conformation when the other molecule is bound. Other proteins appear to resemble elastin, in so far as their function requires that they remain largely unstructured. For example, the abundant nucleoporins that coat the inner surface of the nuclear pore complex form a random coil meshwork that is intimately involved in nuclear transport (see Figure 12–10). Finally, as will be discussed later in this chapter (see Figure 3–80C), unstructured regions of polypeptide chain are often used to connect the binding sites for proteins that function together to catalyze a biological reaction. Thus, for example, in facilitating cell signaling, large *scaffold proteins* use such flexible regions as "tethers" that concentrate sets of interacting proteins, often confining them to particular sites in the cell (discussed in Chapter 15).

We can recognize the unstructured regions in many proteins by their biased amino acid composition: they contain very few of the bulky hydrophobic amino acids that normally form the core of a folded protein, being composed instead of a high proportion of the amino acids Gln, Ser, Pro, Glu, and Lys. Such "natively unfolded" regions also frequently contain repeated sequences of amino acids.

## Covalent Cross-Linkages Often Stabilize Extracellular Proteins

Many protein molecules are either attached to the outside of a cell's plasma membrane or secreted as part of the extracellular matrix. All such proteins are directly exposed to extracellular conditions. To help maintain their structures, the polypeptide chains in such proteins are often stabilized by covalent cross-linkages. These linkages can either tie two amino acids in the same protein together, or connect different polypeptide chains in a multisubunit protein. The most common cross-linkages in proteins are covalent sulfur–sulfur bonds. These *disulfide bonds* (also called *S–S bonds*) form as cells prepare newly synthesized proteins for export. As described in Chapter 12, their formation is catalyzed in the endoplasmic reticulum by an enzyme that links together two pairs of –SH groups of cysteine side chains that are adjacent in the folded protein (**Figure 3–28**). Disulfide bonds do not change the conformation of a protein but instead act as atomic staples to reinforce its most favored conformation. For

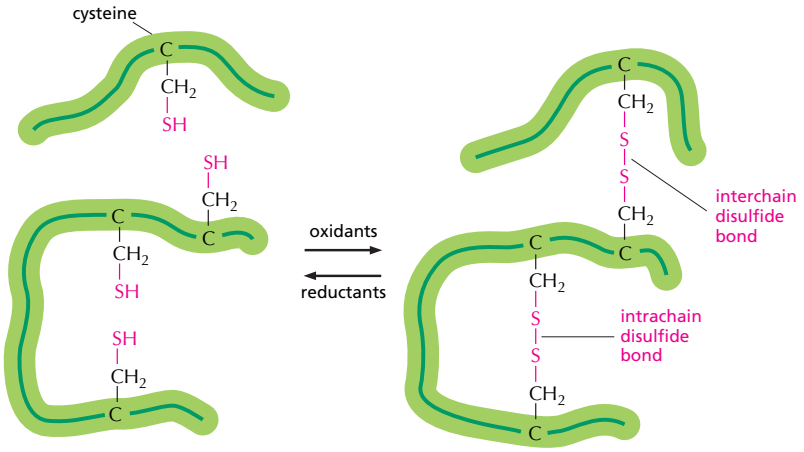



**Figure 3–28 Disulfide bonds.** <ATAC> This diagram illustrates how covalent disulfide bonds form between adjacent cysteine side chains. As indicated, these cross-linkages can join either two parts of the same polypeptide chain or two different polypeptide chains. Since the energy required to break one covalent bond is much larger than the energy required to break even a whole set of noncovalent bonds (see Table 2–1, p. 53), a disulfide bond can have a major stabilizing effect on a protein.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 29 of 240

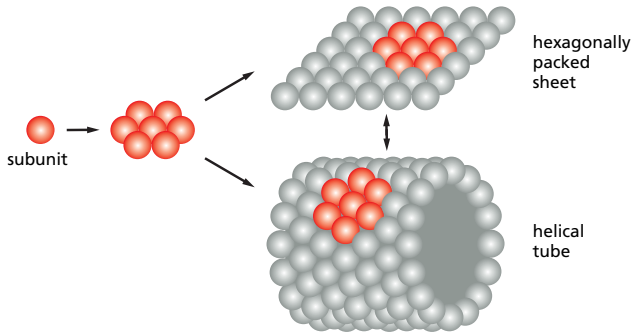



**Figure 3–29 An example of the assembly of a single protein subunit requiring multiple protein–protein contacts.** Hexagonally packed globular protein subunits can form either a flat sheet or a tube.

example, lysozyme—an enzyme in tears that dissolves bacterial cell walls—retains its antibacterial activity for a long time because it is stabilized by such cross-linkages.

Disulfide bonds generally fail to form in the cell cytosol, where a high concentration of reducing agents converts S–S bonds back to cysteine –SH groups. Apparently, proteins do not require this type of reinforcement in the relatively mild environment inside the cell.

## Protein Molecules Often Serve as Subunits for the Assembly of Large Structures

The same principles that enable a protein molecule to associate with itself to form rings or filaments also operate to generate much larger structures in the cell—supramolecular structures such as enzyme complexes, ribosomes, protein filaments, viruses, and membranes. These large objects are not made as single, giant, covalently linked molecules. Instead they are formed by the noncovalent assembly of many separately manufactured molecules, which serve as the subunits of the final structure.

The use of smaller subunits to build larger structures has several advantages:

1. A large structure built from one or a few repeating smaller subunits requires only a small amount of genetic information.
2. Both assembly and disassembly can be readily controlled, reversible processes, because the subunits associate through multiple bonds of relatively low energy.
3. Errors in the synthesis of the structure can be more easily avoided, since correction mechanisms can operate during the course of assembly to exclude malformed subunits.

Some protein subunits assemble into flat sheets in which the subunits are arranged in hexagonal patterns. Specialized membrane proteins are sometimes arranged this way in lipid bilayers. With a slight change in the geometry of the individual subunits, a hexagonal sheet can be converted into a tube (**Figure 3–29**) or, with more changes, into a hollow sphere. Protein tubes and spheres that bind specific RNA and DNA molecules in their interior form the coats of viruses.

The formation of closed structures, such as rings, tubes, or spheres, provides additional stability because it increases the number of bonds between the protein subunits. Moreover, because such a structure is created by mutually dependent, cooperative interactions between subunits, a relatively small change that affects each subunit individually can cause the structure to assemble or disassemble. These principles are dramatically illustrated in the protein coat or *capsid* of many simple viruses, which takes the form of a hollow sphere based on an icosahedron (**Figure 3–30**). Capsids are often made of hundreds of identical protein subunits that enclose and protect the viral nucleic acid (**Figure 3–31**). The protein in such a capsid must have a particularly adaptable structure: not only must it make several different kinds of contacts to create the sphere, it must also change this arrangement to let the nucleic acid out to initiate viral replication once the virus has entered a cell.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 30 of 240

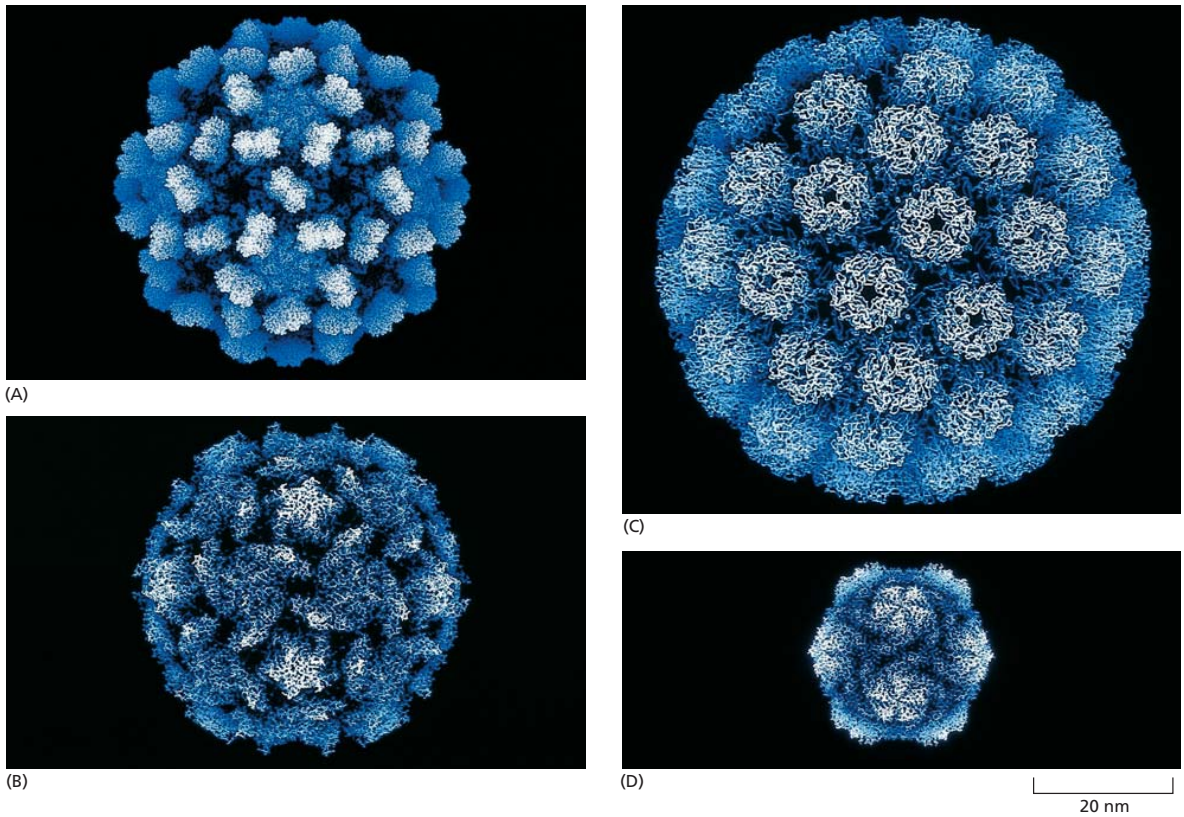



**Figure 3–30 The capsids of some viruses, all shown at the same scale.** (A) Tomato bushy stunt virus; (B) poliovirus; (C) simian virus 40 (SV40); (D) satellite tobacco necrosis virus. The structures of all of these capsids have been determined by x-ray crystallography and are known in atomic detail. (Courtesy of Robert Grant, Stephan Crainic, and James M. Hogle.)

## Many Structures in Cells Are Capable of Self-Assembly

The information for forming many of the complex assemblies of macromolecules in cells must be contained in the subunits themselves, because purified subunits can spontaneously assemble into the final structure under the appropriate conditions. The first large macromolecular aggregate shown to be capable of self-assembly from its component parts was *tobacco mosaic virus (TMV)*. This virus is a long rod in which a cylinder of protein is arranged around a helical RNA core (**Figure 3–32**). If the dissociated RNA and protein subunits are mixed together in solution, they recombine to form fully active viral particles. The assembly process is unexpectedly complex and includes the formation of double rings of protein, which serve as intermediates that add to the growing viral coat.

Another complex macromolecular aggregate that can reassemble from its component parts is the bacterial ribosome. This structure is composed of about 55 different protein molecules and 3 different rRNA molecules. Incubating the individual components under appropriate conditions in a test tube causes them to spontaneously re-form the original structure. Most importantly, such reconstituted ribosomes are able to catalyze protein synthesis. As might be expected, the reassembly of ribosomes follows a specific pathway: after certain proteins have bound to the RNA, this complex is then recognized by other proteins, and so on, until the structure is complete.

It is still not clear how some of the more elaborate self-assembly processes are regulated. Many structures in the cell, for example, seem to have a precisely

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 31 of 240

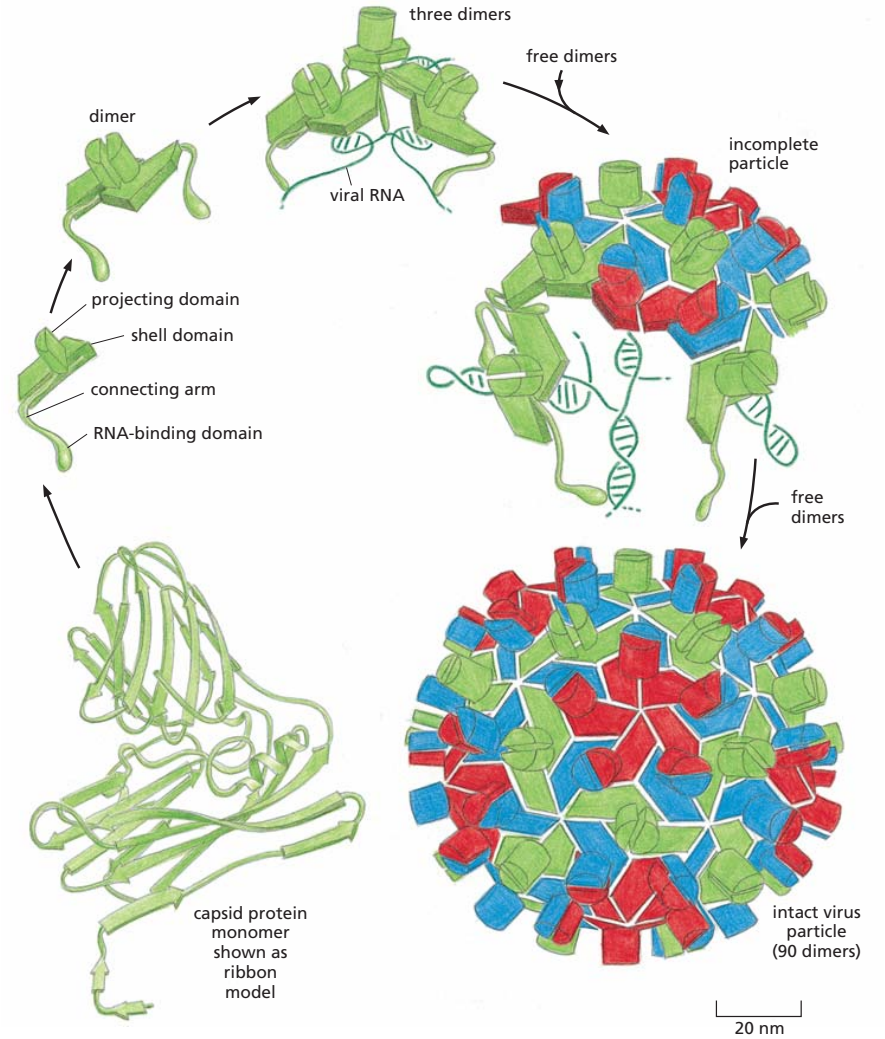
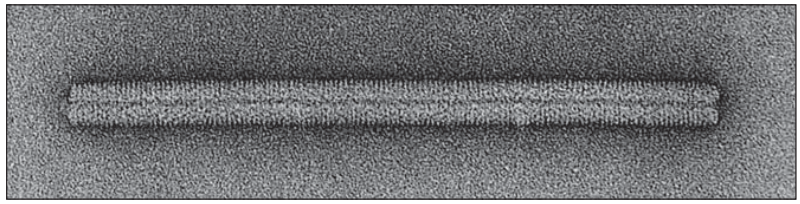


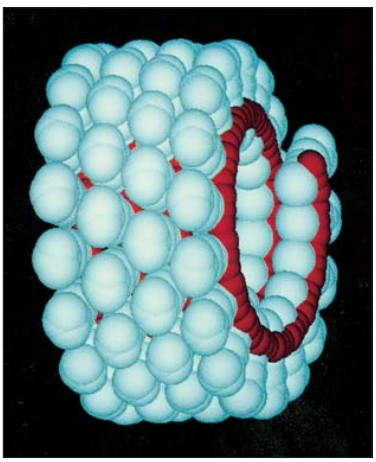

**Figure 3–31 The structure of a spherical virus.** In many viruses, identical protein subunits pack together to create a spherical shell (a capsid) that encloses the viral genome, composed of either RNA or DNA (see also Figure 3–30). For geometric reasons, no more than 60 identical subunits can pack together in a precisely symmetric way. If slight irregularities are allowed, however, more subunits can be used to produce a larger capsid that retains icosahedral symmetry. The tomato bushy stunt virus (TBSV) shown here, for example, is a spherical virus about 33 nm in diameter formed from 180 identical copies of a 386 amino acid capsid protein plus an RNA genome of 4500 nucleotides. To construct such a large capsid, the protein must be able to fit into three somewhat different environments, each of which is differently colored in the virus particle shown here. The postulated pathway of assembly is shown; the precise three-dimensional structure has been determined by x-ray diffraction. (Courtesy of Steve Harrison.)

defined length that is many times greater than that of their component macromolecules. How such length determination is achieved is in many cases a mystery. Three possible mechanisms are illustrated in **Figure 3–33**. In the simplest

**Figure 3–32 The structure of tobacco mosaic virus (TMV).** (A) An electron micrograph of the viral particle, which consists of a single long RNA molecule enclosed in a cylindrical protein coat composed of identical protein subunits. (B) A model showing part of the structure of TMV. A single-stranded RNA molecule of 6395 nucleotides is packaged in a helical coat constructed from 2130 copies of a coat protein 158 amino acids long. Fully infective viral particles can self-assemble in a test tube from purified RNA and protein molecules. (A, courtesy of Robley Williams; B, courtesy of Richard J. Feldmann.)




(A)



(B)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 32 of 240

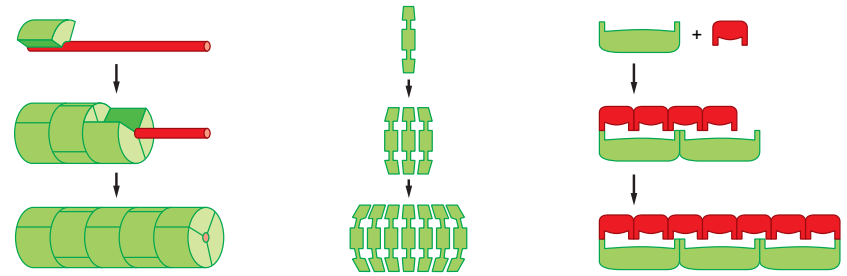



**Figure 3–33 Three mechanisms of length determination for large protein assemblies.** (A) Coassembly along an elongated core protein or other macromolecule that acts as a measuring device. (B) Termination of assembly because of strain that accumulates in the polymeric structure as additional subunits are added, so that beyond a certain length the energy required to fit another subunit onto the chain becomes excessively large. (C) A vernier type of assembly, in which two sets of rodlike molecules differing in length form a staggered complex that grows until their ends exactly match. The name derives from a measuring device based on the same principle, used in mechanical instruments.

case, a long core protein or other macromolecule provides a scaffold that determines the extent of the final assembly. This is the mechanism that determines the length of the TMV particle, where the RNA chain provides the core. Similarly, a core protein is thought to determine the length of the thin filaments in muscle, as well as the length of the long tails of some bacterial viruses (**Figure 3–34**).

## Assembly Factors Often Aid the Formation of Complex Biological Structures

Not all cellular structures held together by noncovalent bonds self-assemble. A mitochondrion, a cilium, or a myofibril of a muscle cell, for example, cannot form spontaneously from a solution of its component macromolecules. In these cases, part of the assembly information is provided by special enzymes and other proteins that perform the function of templates, guiding construction but taking no part in the final assembled structure.

Even relatively simple structures may lack some of the ingredients necessary for their own assembly. In the formation of certain bacterial viruses, for example, the head, which is composed of many copies of a single protein subunit, is assembled on a temporary scaffold composed of a second protein. Because the second protein is absent from the final viral particle, the head structure cannot spontaneously reassemble once it has been taken apart. Other examples are known in which proteolytic cleavage is an essential and irreversible step in the normal assembly process. This is even the case for some small protein assemblies, including the structural protein collagen and the hormone insulin (**Figure 3–35**). From these relatively simple examples, it seems certain that the assembly of a structure as complex as a mitochondrion or a cilium will involve temporal and spatial ordering imparted by numerous other cell components.




**Figure 3–34 An electron micrograph of bacteriophage lambda.** The tip of the virus tail attaches to a specific protein on the surface of a bacterial cell, after which the tightly packaged DNA in the head is injected through the tail into the cell. The tail has a precise length, determined by the mechanism shown in Figure 3–33A.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 33 of 240

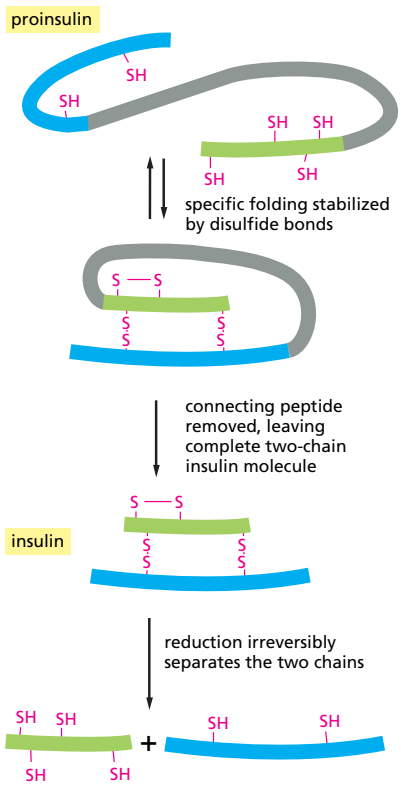



**Figure 3–35 Proteolytic cleavage in insulin assembly.** The polypeptide hormone insulin cannot spontaneously re-form efficiently if its disulfide bonds are disrupted. It is synthesized as a larger protein *(proinsulin)* that is cleaved by a proteolytic enzyme after the protein chain has folded into a specific shape. Excision of part of the proinsulin polypeptide chain removes some of the information needed for the protein to fold spontaneously into its normal conformation. Once insulin has been denatured and its two polypeptide chains have separated, its ability to reassemble is lost.

## Summary

*A protein molecule's amino acid sequence determines its three-dimensional conformation. Noncovalent interactions between different parts of the polypeptide chain stabilize its folded structure. The amino acids with hydrophobic side chains tend to cluster in the interior of the molecule, and local hydrogen-bond interactions between neighboring peptide bonds give rise to α helices and β sheets.*

*Globular regions, known as domains, are the modular units from which many proteins are constructed; such domains generally contain 40–350 amino acids. Small proteins typically consist of only a single domain, while large proteins are formed from several domains linked together by various lengths of polypeptide chain, some of which can be relatively disordered. As proteins have evolved, domains have been modified and combined with other domains to construct new proteins. Thus far, about 800 different ways of folding up a domain have been observed, among more than 20,000 known protein structures.*

*Proteins are brought together into larger structures by the same noncovalent forces that determine protein folding. Proteins with binding sites for their own surface can assemble into dimers, closed rings, spherical shells, or helical polymers. Although mixtures of proteins and nucleic acids can assemble spontaneously into complex structures in a test tube, many biological assembly processes involve irreversible steps. Consequently, not all structures in the cell are capable of spontaneous reassembly after they have been dissociated into their component parts.*

# PROTEIN FUNCTION

We have seen that each type of protein consists of a precise sequence of amino acids that allows it to fold up into a particular three-dimensional shape, or conformation. But proteins are not rigid lumps of material. They often have precisely engineered moving parts whose mechanical actions are coupled to chemical events. It is this coupling of chemistry and movement that gives proteins the extraordinary capabilities that underlie the dynamic processes in living cells.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 34 of 240

In this section, we explain how proteins bind to other selected molecules and how their activity depends on such binding. We show that the ability to bind to other molecules enables proteins to act as catalysts, signal receptors, switches, motors, or tiny pumps. The examples we discuss in this chapter by no means exhaust the vast functional repertoire of proteins. You will encounter the specialized functions of many other proteins elsewhere in this book, based on similar principles.

## All Proteins Bind to Other Molecules

A protein molecule's physical interaction with other molecules determines its biological properties. Thus, antibodies attach to viruses or bacteria to mark them for destruction, the enzyme hexokinase binds glucose and ATP so as to catalyze a reaction between them, actin molecules bind to each other to assemble into actin filaments, and so on. Indeed, all proteins stick, or *bind*, to other molecules. In some cases, this binding is very tight; in others it is weak and short-lived. But the binding always shows great *specificity*, in the sense that each protein molecule can usually bind just one or a few molecules out of the many thousands of different types it encounters. The substance that is bound by the protein—whether it is an ion, a small molecule, or a macromolecule such as another protein—is referred to as a **ligand** for that protein (from the Latin word *ligare*, meaning "to bind").

The ability of a protein to bind selectively and with high affinity to a ligand depends on the formation of a set of weak, noncovalent bonds—hydrogen bonds, electrostatic attractions, and van der Waals attractions—plus favorable hydrophobic interactions (see Panel 2–3, pp. 110–111). Because each individual bond is weak, effective binding occurs only when many of these bonds form simultaneously. Such binding is possible only if the surface contours of the ligand molecule fit very closely to the protein, matching it like a hand in a glove (**Figure 3–36**).

The region of a protein that associates with a ligand, known as the ligand's *binding site,* usually consists of a cavity in the protein surface formed by a particular arrangement of amino acids. These amino acids can belong to different portions of the polypeptide chain that are brought together when the protein folds (**Figure 3–37**). Separate regions of the protein surface generally provide binding sites for different ligands, allowing the protein's activity to be regulated, as we shall see later. And other parts of the protein act as a handle to position the protein in the cell—an example is the SH2 domain discussed previously, which often moves a protein containing it to particular intracellular sites in response to particular signals.

Although the atoms buried in the interior of the protein have no direct contact with the ligand, they form the framework that gives the surface its contours and its chemical and mechanical properties. Even small changes to the amino acids in the interior of a protein molecule can change its three-dimensional shape enough to destroy a binding site on the surface.

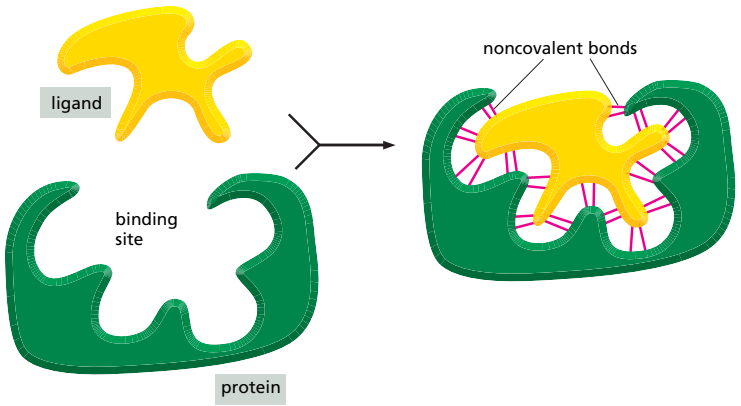



**Figure 3–36 The selective binding of a protein to another molecule.** Many weak bonds are needed to enable a protein to bind tightly to a second molecule, which is called a *ligand* for the protein. A ligand must therefore fit precisely into a protein's binding site, like a hand into a glove, so that a large number of noncovalent bonds form between the protein and the ligand.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 35 of 240

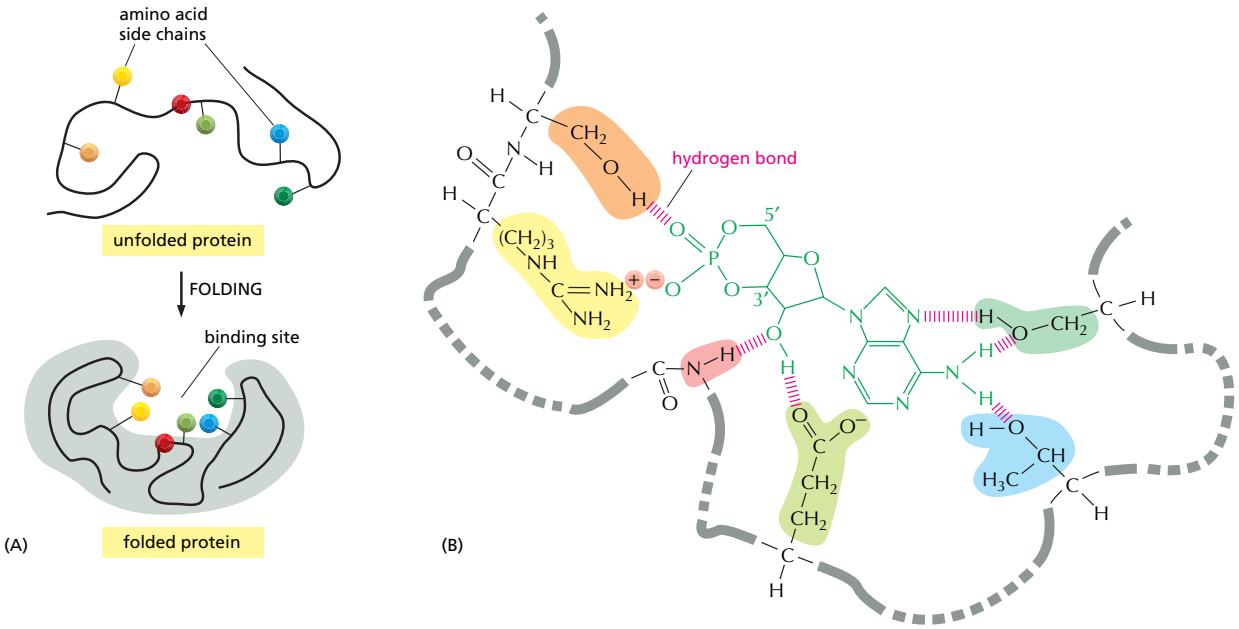



**Figure 3–37 The binding site of a protein.** (A) The folding of the polypeptide chain typically creates a crevice or cavity on the protein surface. This crevice contains a set of amino acid side chains disposed in such a way that they can form noncovalent bonds only with certain ligands. (B) A close-up of an actual binding site showing the hydrogen bonds and electrostatic interactions formed between a protein and its ligand. In this example, cyclic AMP is the bound ligand.

## The Surface Conformation of a Protein Determines Its Chemistry

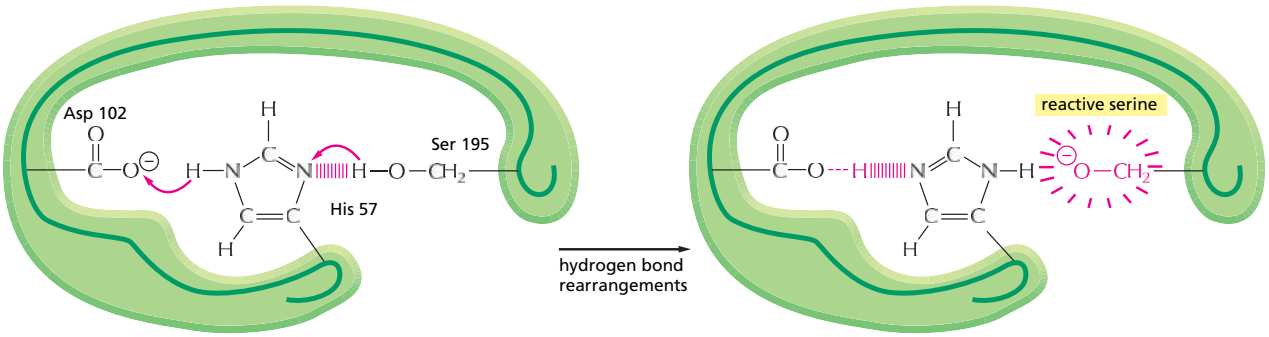
Proteins have impressive chemical capabilities because the neighboring chemical groups on their surface often interact in ways that enhance the chemical reactivity of amino acid side chains. These interactions fall into two main categories.

First, the interaction of neighboring parts of the polypeptide chain may restrict the access of water molecules to that protein's ligand-binding sites. This is important because water molecules readily form hydrogen bonds that can compete with ligands for sites on the protein surface. Proteins and their ligands form tighter hydrogen bonds (and electrostatic interactions) if the protein can exclude water molecules from its binding sites. It might be hard to imagine a mechanism that would exclude a molecule as small as water from a protein surface without affecting the access of the ligand itself. However, because of the strong tendency of water molecules to form water–water hydrogen bonds, water molecules exist in a large hydrogen-bonded network (see Panel 2–2, pp. 108–109). In effect, a protein can keep a ligand-binding site dry because it is energetically unfavorable for individual water molecules to break away from this network, as they must do to reach into a crevice on a protein's surface.

Second, the clustering of neighboring polar amino acid side chains can alter their reactivity. If protein folding forces together a number of negatively charged side chains against their mutual repulsion, for example, the affinity of the site for a positively charged ion is greatly increased. In addition, when amino acid side chains interact with one another through hydrogen bonds, normally unreactive side groups (such as the $–CH_2OH$ on the serine shown in **Figure 3–38**) can become reactive, enabling them to be used to make or break selected covalent bonds.

The surface of each protein molecule therefore has a unique chemical reactivity that depends not only on which amino acid side chains are exposed, but

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 36 of 240




also on their exact orientation relative to one another. For this reason, two even slightly different conformations of the same protein molecule may differ greatly in their chemistry.

**Figure 3–38 An unusually reactive amino acid at the active site of an enzyme.** This example is the "catalytic triad" found in chymotrypsin, elastase, and other serine proteases (see Figure 3–12). The aspartic acid side chain (Asp 102) induces the histidine (His 57) to remove the proton from serine 195. This activates the serine to form a covalent bond with the enzyme substrate, hydrolyzing a peptide bond. The many convolutions of the polypeptide chain are omitted here.

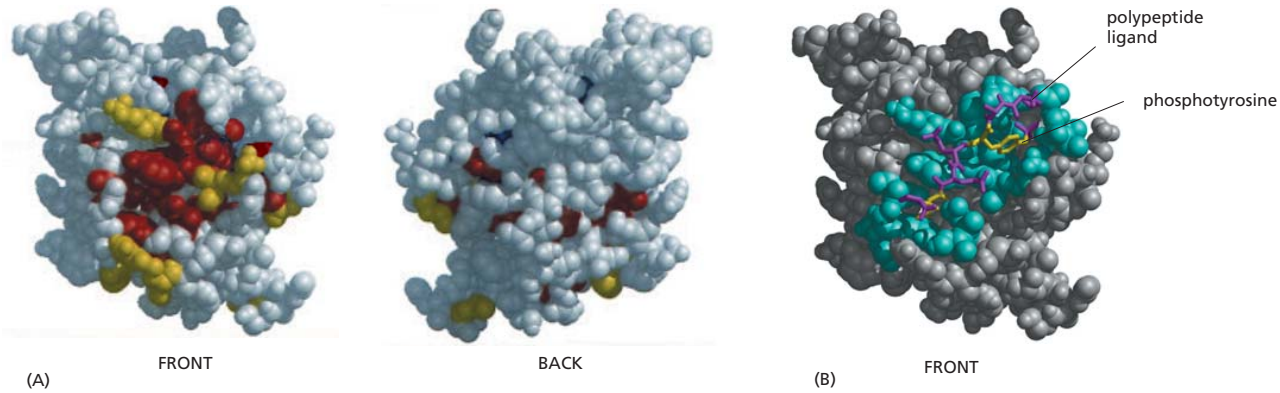
## Sequence Comparisons Between Protein Family Members Highlight Crucial Ligand-Binding Sites

As we have described previously, genome sequences allow us to group many of the domains in proteins into families that show clear evidence of their evolution from a common ancestor. The three-dimensional structures of the members of the same domain family are remarkably similar. For example, even when the amino acid sequence identity falls to 25%, the backbone atoms in a domain follow a common protein fold within 0.2 nanometers (2 Å).

We can therefore use a method called evolutionary tracing to identify those sites in a protein domain that are the most crucial to the domain's function. For this purpose, those amino acids that are unchanged, or nearly unchanged, in all of the known protein family members are mapped onto a model of the three-dimensional structure of one family member. When this is done, the most invariant positions often form one or more clusters on the protein surface, as illustrated in **Figure 3–39**A for the SH2 domain described previously (see Panel 3–2, pp. 132–133). These clusters generally correspond to ligand binding sites.

The SH2 domain is a module that functions in protein–protein interactions. It binds the protein containing it to a second protein that contains a phosphorylated tyrosine side chain in a specific amino acid sequence context, as shown in Figure 3–39B. The amino acids located at the binding site for the phosphorylated polypeptide have been the slowest to change during the long evolutionary process that produced the large SH2 family of peptide recognition domains. Because mutation is a random process, this result is attributed to the preferential elimination during evolution of all organisms whose SH2 domains became altered in a way that inactivated the SH2-binding site, thereby destroying the function of the SH2 domain.

**Figure 3–39 The evolutionary trace method applied to the SH2 domain.** (A) Front and back views of a space-filling model of the SH2 domain, with evolutionarily conserved amino acids on the protein surface colored *yellow*, and those more toward the protein interior colored *red*. (B) The structure of the SH2 domain with its bound polypeptide. Here, those amino acids located within 0.4 nm of the bound ligand are colored *blue*. The two key amino acids of the ligand are *yellow*, and the others are *purple*. Note the high degree of correspondance between (A) and (B). (Adapted from O. Lichtarge, H.R. Bourne and F.E. Cohen, *J. Mol. Biol.* 257:342–358, 1996. With permission from Elsevier.)




LCY Biotechnology Holding, Inc.
Ex. 1008
Page 37 of 240

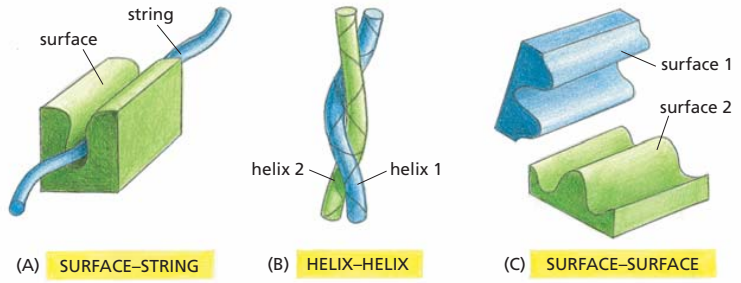



**Figure 3–40 Three ways in which two proteins can bind to each other.** Only the interacting parts of the two proteins are shown. (A) A rigid surface on one protein can bind to an extended loop of polypeptide chain (a "string") on a second protein. (B) Two α helices can bind together to form a coiled-coil. (C) Two complementary rigid surfaces often link two proteins together.

In this era of extensive genome sequencing, many new protein families have been discovered whose functions are unknown. Once a three-dimensional structure has been determined for one family member, evolutionary tracing allows biologists to determine binding sites for the members of that family, thereby helping to decipher protein function.

## Proteins Bind to Other Proteins Through Several Types of Interfaces

Proteins can bind to other proteins in at least three ways. In many cases, a portion of the surface of one protein contacts an extended loop of polypeptide chain (a "string") on a second protein (**Figure 3–40**A). Such a surface–string interaction, for example, allows the SH2 domain to recognize a phosphorylated polypeptide loop on a second protein, as just described, and it also enables a protein kinase to recognize the proteins that it will phosphorylate (see below).

A second type of protein–protein interface forms when two α helices, one from each protein, pair together to form a coiled-coil (Figure 3–40B). This type of protein interface is found in several families of gene regulatory proteins, as discussed in Chapter 7.

The most common way for proteins to interact, however, is by the precise matching of one rigid surface with that of another (Figure 3–40C). Such interactions can be very tight, since a large number of weak bonds can form between two surfaces that match well. For the same reason, such surface–surface interactions can be extremely specific, enabling a protein to select just one partner from the many thousands of different proteins found in a cell.

## Antibody Binding Sites Are Especially Versatile <GCCG>

All proteins must bind to particular ligands to carry out their various functions. The antibody family is notable for its capacity for tight selective binding (discussed in detail in Chapter 25).

**Antibodies**, or immunoglobulins, are proteins produced by the immune system in response to foreign molecules, such as those on the surface of an invading microorganism. Each antibody binds tightly to a particular target molecule, thereby either inactivating the target molecule directly or marking it for destruction. An antibody recognizes its target (called an **antigen**) with remarkable specificity. Because there are potentially billions of different antigens that humans might encounter, we have to be able to produce billions of different antibodies.

Antibodies are Y-shaped molecules with two identical binding sites that are complementary to a small portion of the surface of the antigen molecule. A detailed examination of the antigen-binding sites of antibodies reveals that they are formed from several loops of polypeptide chain that protrude from the ends of a pair of closely juxtaposed protein domains (**Figure 3–41**). Different antibodies generate an enormous diversity of antigen-binding sites by changing only the length and amino acid sequence of these loops, without altering the basic protein structure.

Loops of this kind are ideal for grasping other molecules. They allow a large number of chemical groups to surround a ligand so that the protein can link to

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 38 of 240

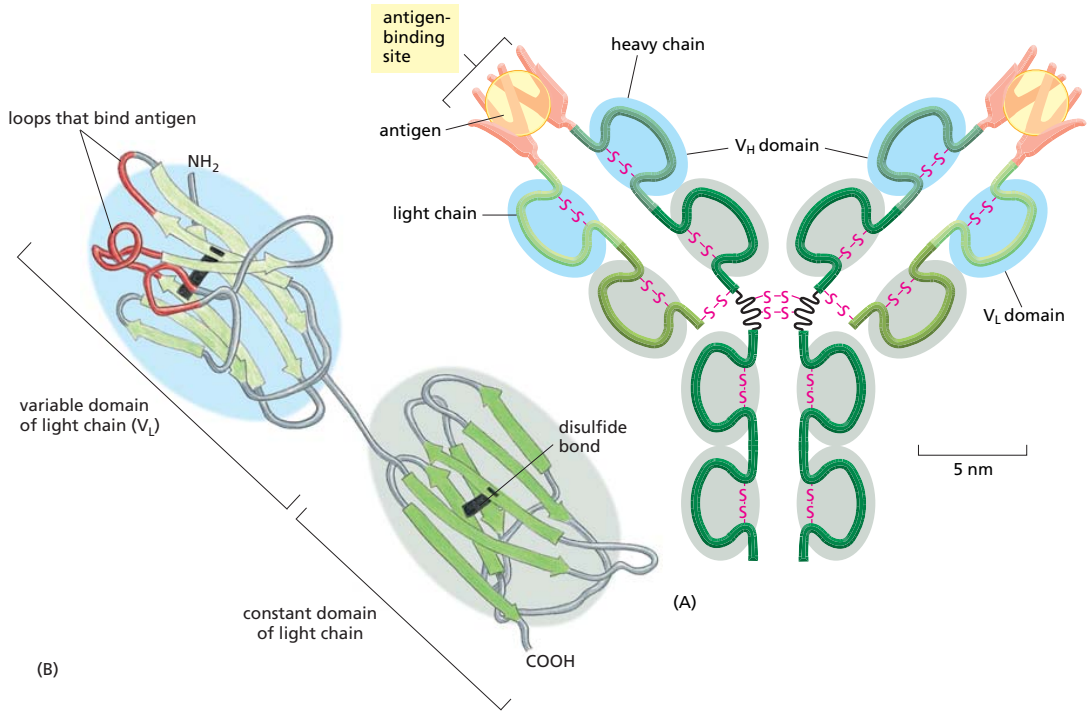



**Figure 3–41 An antibody molecule.** (A) A typical antibody molecule is Y-shaped and has two identical binding sites for its antigen, one on each arm of the Y. The protein is composed of four polypeptide chains (two identical heavy chains and two identical and smaller light chains) held together by disulfide bonds. Each chain is made up of several different immunoglobulin domains, here shaded either *blue* or *gray*. The antigen-binding site is formed where a heavy-chain variable domain ($V_H$) and a light-chain variable domain ($V_L$) come close together. These are the domains that differ most in their sequence and structure in different antibodies. Each domain at the end of the two arms of the antibody molecule forms loops that bind to the antigen. In (B) we can see these fingerlike loops *(red)* contributed by the $V_L$ domain.

it with many weak bonds. For this reason, loops often form the ligand-binding sites in proteins.

## The Equilibrium Constant Measures Binding Strength

Molecules in the cell encounter each other very frequently because of their continual random thermal movements. Colliding molecules with poorly matching surfaces form few noncovalent bonds with one another, and the two molecules dissociate as rapidly as they come together. At the other extreme, when many noncovalent bonds form between two colliding molecules, the association can persist for a very long time (**Figure 3–42**). Strong interactions occur in cells whenever a biological function requires that molecules remain associated for a long time—for example, when a group of RNA and protein molecules come together to make a subcellular structure such as a ribosome.

We can measure the strength with which any two molecules bind to each other. As an example, consider a population of identical antibody molecules that suddenly encounters a population of ligands diffusing in the fluid surrounding them. At frequent intervals, one of the ligand molecules will bump into the binding site of an antibody and form an antibody–ligand complex. The population of antibody–ligand complexes will therefore increase, but not without limit: over time, a second process, in which individual complexes break apart because of thermally induced motion, will become increasingly important. Eventually, any population of antibody molecules and ligands will reach a steady state, or equilibrium, in which the number of binding (association) events per second is precisely equal to the number of "unbinding" (dissociation) events (see Figure 2–52).

From the concentrations of the ligand, antibody, and antibody–ligand complex at equilibrium, we can calculate a convenient measure—the **equilibrium constant** ***(K)***—of the strength of binding (**Figure 3–43**A). The equilibrium constant for a reaction in which two molecules (A and B) bind to each other to form a complex (AB) has units of liters/mole, and half of the binding sites will be occupied by ligand when that ligand's concentration (in moles/liter) reaches a value that is equal to $1/K$. This equilibrium constant is larger the greater the binding strength, and it is a direct measure of the free-energy difference

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 39 of 240

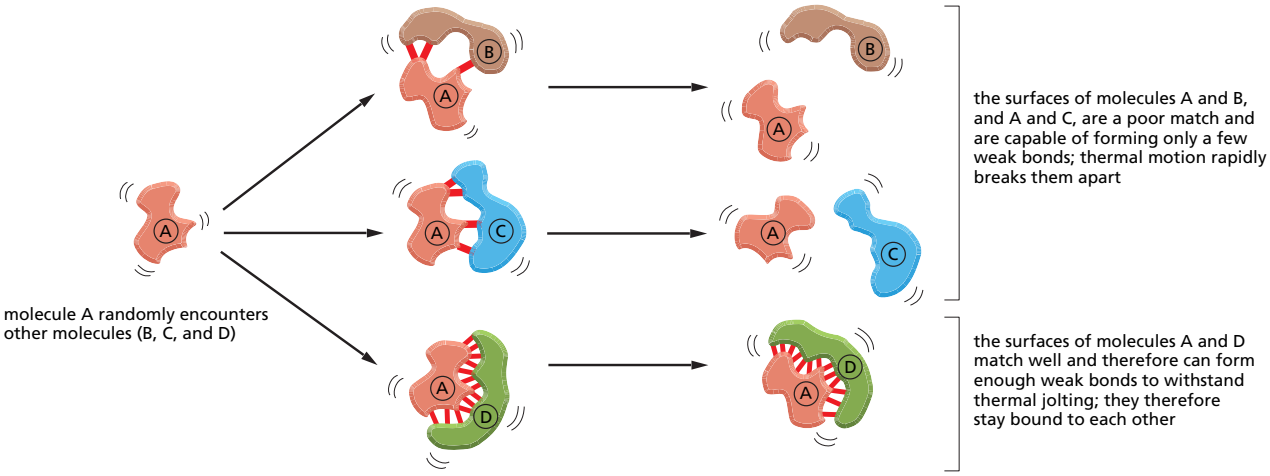



**Figure 3–42 How noncovalent bonds mediate interactions between macromolecules.**

between the bound and free states (Figure 3–43B and C). Even a change of a few noncovalent bonds can have a striking effect on a binding interaction, as shown by the example in **Figure 3–44**. (Note that the equilibrium constant, as defined here is also known as the association or affinity constant, $K_a$.)

We have used the case of an antibody binding to its ligand to illustrate the effect of binding strength on the equilibrium state, but the same principles apply to any molecule and its ligand. Many proteins are enzymes, which, as we now discuss, first bind to their ligands and then catalyze the breakage or formation of covalent bonds in these molecules.

## Enzymes Are Powerful and Highly Specific Catalysts

Many proteins can perform their function simply by binding to another molecule. An actin molecule, for example, need only associate with other actin

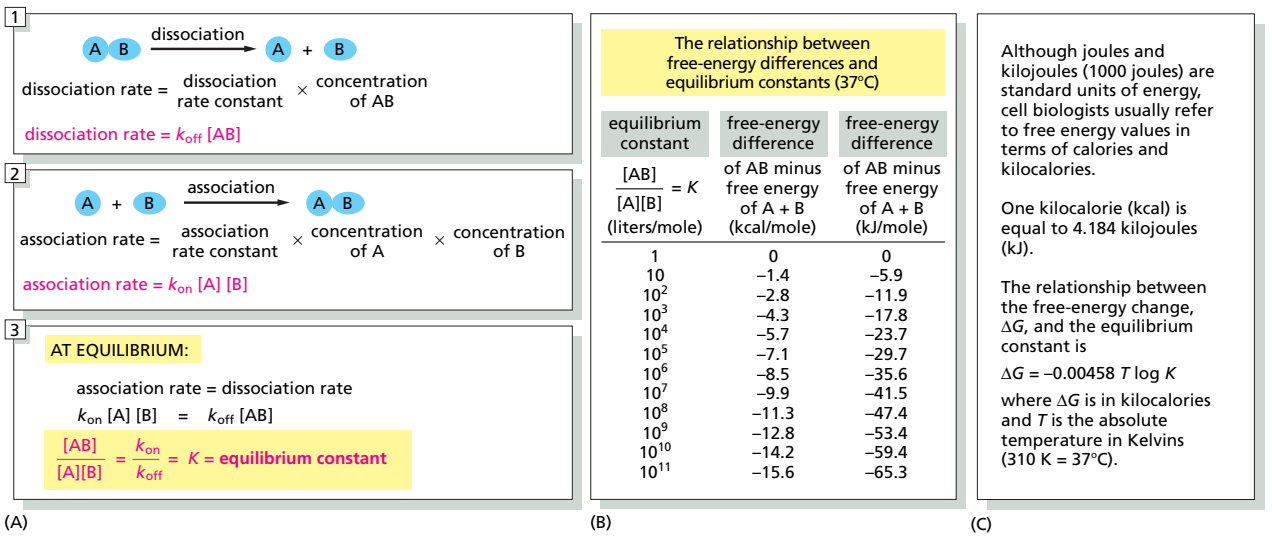



**Figure 3–43 Relating binding energies to the equilibrium constant for an association reaction.** (A) The equilibrium between molecules A and B and the complex AB is maintained by a balance between the two opposing reactions shown in panels 1 and 2. Molecules A and B must collide if they are to react, and the association rate is therefore proportional to the product of their individual concentrations [A] × [B]. (Square brackets indicate concentration.) As shown in panel 3, the ratio of the rate constants for the association and the dissociation reactions is equal to the equilibrium constant (*K*) for the reaction. (B) The equilibrium constant in panel 3 is that for the reaction A + B ⇌ AB, and the larger its value, the stronger the binding between A and B. Note that for every 1.41 kcal/mole (5.91 kJ/mole) decrease in free energy the equilibrium constant increases by a factor of 10 at 37°C.

The equilibrium constant here has units of liters/mole: for simple binding interactions it is also called the *affinity constant* or *association constant,* denoted $K_a$. The reciprocal of $K_a$ is called the dissociation constant, $K_d$ (in units of moles/liter).

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 40 of 240

molecules to form a filament. There are other proteins, however, for which ligand binding is only a necessary first step in their function. This is the case for the large and very important class of proteins called **enzymes**. As described in Chapter 2, enzymes are remarkable molecules that determine all the chemical transformations that make and break covalent bonds in cells. They bind to one or more ligands, called **substrates**, and convert them into one or more chemically modified *products*, doing this over and over again with amazing rapidity. Enzymes speed up reactions, often by a factor of a million or more, without themselves being changed—that is, they act as **catalysts** that permit cells to make or break covalent bonds in a controlled way. It is the catalysis of organized sets of chemical reactions by enzymes that creates and maintains the cell, making life possible.

We can group enzymes into functional classes that perform similar chemical reactions (**Table 3–1**). Each type of enzyme within such a class is highly specific, catalyzing only a single type of reaction. Thus, *hexokinase* adds a phosphate group to D-glucose but ignores its optical isomer L-glucose; the blood-clotting enzyme *thrombin* cuts one type of blood protein between a particular arginine and its adjacent glycine and nowhere else, and so on. As discussed in detail in Chapter 2, enzymes work in teams, with the product of one enzyme becoming the substrate for the next. The result is an elaborate network of metabolic pathways that provides the cell with energy and generates the many large and small molecules that the cell needs (see Figure 2–35).

Consider 1000 molecules of A and 1000 molecules of B in a eucaryotic cell. The concentration of both will be about $10^{-9}$ M.

If the equilibrium constant ($K$) for A + B ⇌ AB is $10^{10}$, then one can calculate that at equilibrium there will be

| 270 | 270 | 730 |
|---|---|---|
| A molecules | B molecules | AB molecules |

If the equilibrium constant is a little weaker at $10^{8}$, which represents a loss of 2.8 kcal/mole of binding energy from the example above, or 2–3 fewer hydrogen bonds, then there will be

| 915 | 915 | 85 |
|---|---|---|
| A molecules | B molecules | AB molecules |

**Figure 3–44 Small changes in the number of weak bonds can have drastic effects on a binding interaction.** This example illustrates the dramatic effect of the presence or absence of a few weak noncovalent bonds in a biological context.

## Substrate Binding Is the First Step in Enzyme Catalysis

For a protein that catalyzes a chemical reaction (an enzyme), the binding of each substrate molecule to the protein is an essential prelude. In the simplest case, if we denote the enzyme by E, the substrate by S, and the product by P, the basic reaction path is E + S → ES → EP → E + P. From this reaction path, we see that there is a limit to the amount of substrate that a single enzyme molecule can process in a given time. An increase in the concentration of substrate also increases the rate at which product is formed, up to a maximum value (**Figure 3–45**). At that point the enzyme molecule is saturated with substrate, and the rate of reaction ($V_{max}$) depends only on how rapidly the enzyme can process the substrate molecule. This maximum rate divided by the enzyme concentration is

**Table 3–1 Some Common Types of Enzymes**

| ENZYME | REACTION CATALYZED |
|---|---|
| Hydrolases | general term for enzymes that catalyze a hydrolytic cleavage reaction; *nucleases* and *proteases* are more specific names for subclasses of these enzymes. |
| Nucleases | break down nucleic acids by hydrolyzing bonds between nucleotides. |
| Proteases | break down proteins by hydrolyzing bonds between amino acids. |
| Synthases | synthesize molecules in anabolic reactions by condensing two smaller molecules together. |
| Isomerases | catalyze the rearrangement of bonds within a single molecule. |
| Polymerases | catalyze polymerization reactions such as the synthesis of DNA and RNA. |
| Kinases | catalyze the addition of phosphate groups to molecules. Protein kinases are an important group of kinases that attach phosphate groups to proteins. |
| Phosphatases | catalyze the hydrolytic removal of a phosphate group from a molecule. |
| Oxido-Reductases | general name for enzymes that catalyze reactions in which one molecule is oxidized while the other is reduced. Enzymes of this type are often more specifically named either *oxidases, reductases,* or *dehydrogenases.* |
| ATPases | hydrolyze ATP. Many proteins with a wide range of roles have an energy-harnessing ATPase activity as part of their function, for example, motor proteins such as *myosin* and membrane transport proteins such as the *sodium–potassium pump.* |

Enzyme names typically end in "-ase," with the exception of some enzymes, such as pepsin, trypsin, thrombin and lysozyme that were discovered and named before the convention became generally accepted at the end of the nineteenth century. The common name of an enzyme usually indicates the substrate and the nature of the reaction catalyzed. For example, citrate synthase catalyzes the synthesis of citrate by a reaction between acetyl CoA and oxaloacetate.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 41 of 240

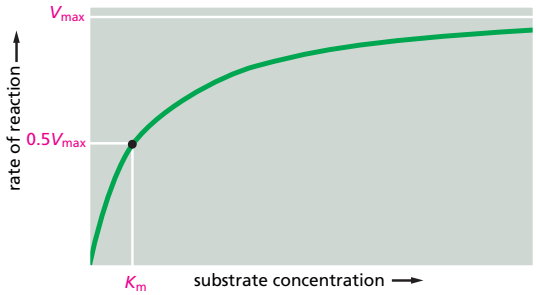



**Figure 3–45 Enzyme kinetics.** The rate of an enzyme reaction *(V)* increases as the substrate concentration increases until a maximum value ($V_{max}$) is reached. At this point all substrate-binding sites on the enzyme molecules are fully occupied, and the rate of reaction is limited by the rate of the catalytic process on the enzyme surface. For most enzymes, the concentration of substrate ($K_m$) at which the reaction rate is half-maximal (*black dot*) is a measure of how tightly the substrate is bound, with a large value of $K_m$ corresponding to weak binding.

called the *turnover number*. The turnover number is often about 1000 substrate molecules processed per second per enzyme molecule, although turnover numbers between 1 and 10,000 are known.

The other kinetic parameter frequently used to characterize an enzyme is its $K_m$, the concentration of substrate that allows the reaction to proceed at one-half its maximum rate (0.5 $V_{max}$) (see Figure 3–45). A *low* $K_m$ value means that the enzyme reaches its maximum catalytic rate at a *low concentration* of substrate and generally indicates that the enzyme binds to its substrate very tightly, whereas a *high* $K_m$ value corresponds to weak binding. The methods used to characterize enzymes in this way are explained in **Panel 3–3** (pp. 162–163).

## Enzymes Speed Reactions by Selectively Stabilizing Transition States

Enzymes achieve extremely high rates of chemical reaction—rates that are far higher than for any synthetic catalysts. There are several reasons for this efficiency. First, the enzyme increases the local concentration of substrate molecules at the catalytic site and holds all the appropriate atoms in the correct orientation for the reaction that is to follow. More importantly, however, some of the binding energy contributes directly to the catalysis. Substrate molecules must pass through a series of intermediate states of altered geometry and electron distribution before they form the ultimate products of the reaction. The free energy required to attain the most unstable **transition state** is called the *activation energy* for the reaction, and it is the major determinant of the reaction rate. Enzymes have a much higher affinity for the transition state of the substrate than they have for the stable form. Because this tight binding greatly lowers the energies of the transition state, the enzyme greatly accelerates a particular reaction by lowering the activation energy that is required (**Figure 3–46**).

By intentionally producing antibodies that act like enzymes, we can demonstrate that stabilizing a transition state can greatly increase a reaction rate. Consider, for example, the hydrolysis of an amide bond, which is similar to the peptide bond that joins two adjacent amino acids in a protein. In an aqueous solution, an amide bond hydrolyzes very slowly by the mechanism shown in **Figure 3–47**A. In the central intermediate, or transition state, the carbonyl carbon is bonded to four atoms arranged at the corners of a tetrahedron. By generating monoclonal antibodies that bind tightly to a stable analog of this very unstable *tetrahedral intermediate*, one can obtain an antibody that functions like an enzyme (Figure 3–47B). Because this *catalytic antibody* binds to and stabilizes the tetrahedral intermediate, it increases the spontaneous rate of amide-bond hydrolysis more than 10,000-fold.

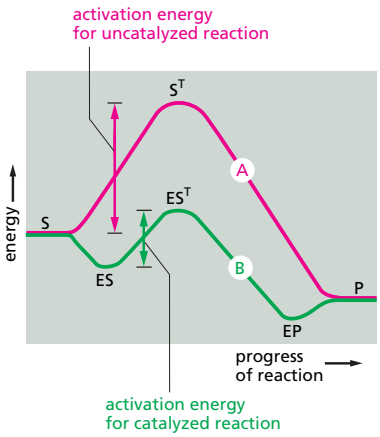



**Figure 3–46 Enzymatic acceleration of chemical reactions by decreasing the activation energy.** Often both the uncatalyzed reaction (A) and the enzyme-catalyzed reaction (B) can go through several transition states. It is the transition state with the highest energy ($S^T$ and $ES^T$) that determines the activation energy and limits the rate of the reaction. (S = substrate; P = product of the reaction; ES = enzyme–substrate complex; EP = enzyme–product complex.)

## Enzymes Can Use Simultaneous Acid and Base Catalysis

**Figure 3–48** compares the spontaneous reaction rates and the corresponding enzyme-catalyzed rates for five enzymes. Rate accelerations range from $10^9$ to $10^{23}$. Clearly, enzymes are much better catalysts than catalytic antibodies.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 42 of 240

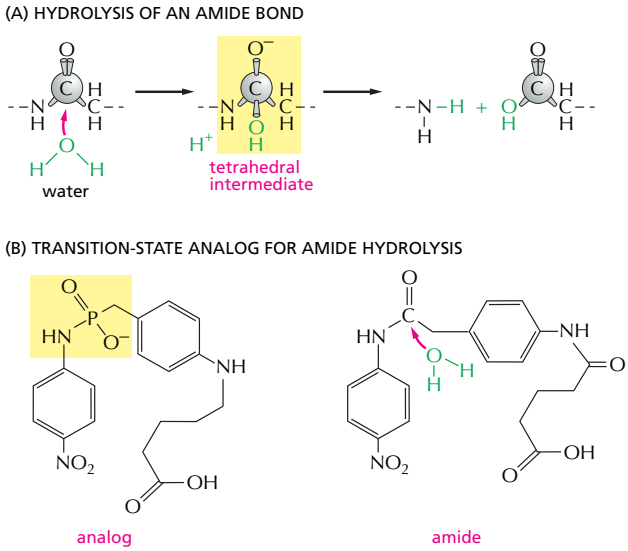



**Figure 3–47 Catalytic antibodies.** The stabilization of a transition state by an antibody creates an enzyme. (A) The reaction path for the hydrolysis of an amide bond goes through a tetrahedral intermediate, the high-energy transition state for the reaction. (B) The molecule on the left was covalently linked to a protein and used as an antigen to generate an antibody that binds tightly to the region of the molecule shown in *yellow*. Because this antibody also bound tightly to the transition state in (A), it was found to function as an enzyme that efficiently catalyzed the hydrolysis of the amide bond in the molecule on the right.

Enzymes not only bind tightly to a transition state, they also contain precisely positioned atoms that alter the electron distributions in those atoms that participate directly in the making and breaking of covalent bonds. Peptide bonds, for example, can be hydrolyzed in the absence of an enzyme by exposing a polypeptide to either a strong acid or a strong base, as illustrated in **Figure 3–49**. Enzymes are unique, however, in being able to use acid and base catalysis simultaneously, since the rigid framework of the protein binds the acidic and basic residues and prevents them from combining with each other (as they would do in solution) (Figure 3–49D).

The fit between an enzyme and its substrate needs to be precise. A small change introduced by genetic engineering in the active site of an enzyme can have a profound effect. Replacing a glutamic acid with an aspartic acid in one enzyme, for example, shifts the position of the catalytic carboxylate ion by only 1 Å (about the radius of a hydrogen atom); yet this is enough to decrease the activity of the enzyme a thousandfold.

## Lysozyme Illustrates How an Enzyme Works <AGCA>

To demonstrate how enzymes catalyze chemical reactions, we examine an enzyme that acts as a natural antibiotic in egg white, saliva, tears, and other secretions. **Lysozyme** catalyzes the cutting of polysaccharide chains in the cell walls of bacteria. Because the bacterial cell is under pressure from osmotic forces, cutting even a small number of polysaccharide chains causes the cell wall to rupture and the cell to burst. Lysozyme is a relatively small and stable protein

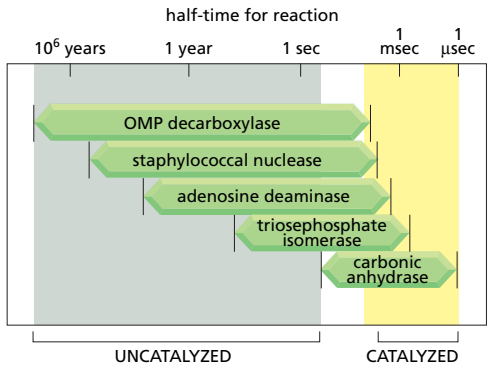



**Figure 3–48 The rate accelerations caused by five different enzymes.** (Adapted from A. Radzicka and R. Wolfenden, *Science* 267:90–93, 1995. With permission from AAAS.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 43 of 240

## WHY ANALYZE THE KINETICS OF ENZYMES?

Enzymes are the most selective and powerful catalysts known. An understanding of their detailed mechanisms provides a critical tool for the discovery of new drugs, for the large-scale industrial synthesis of useful chemicals, and for appreciating the chemistry of cells and organisms. A detailed study of the rates of the chemical reactions that are catalyzed by a purified enzyme—more specifically how these rates change with changes in conditions such as the concentrations of substrates, products, inhibitors, and regulatory ligands—allows biochemists to figure out exactly how each enzyme works. For example, this is the way that the ATP-producing reactions of glycolysis, shown previously in Figure 2–72, were deciphered—allowing us to appreciate the rationale for this critical enzymatic pathway.

In this Panel, we introduce the important field of enzyme kinetics, which has been indispensable for deriving much of the detailed knowledge that we now have about cell chemistry.

## STEADY-STATE ENZYME KINETICS

Many enzymes have only one substrate, which they bind and then process to produce products according to the scheme outlined in Figure 3–50A. In this case, the reaction is written as

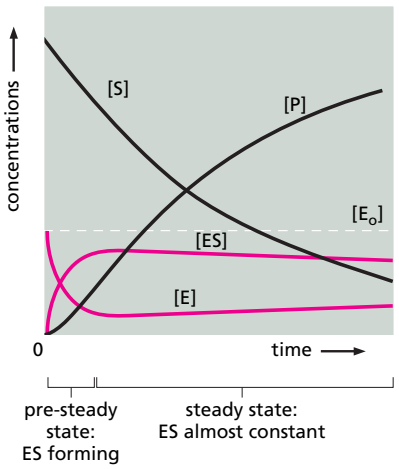
$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_{cat}} E + P$$

Here we have assumed that the reverse reaction, in which E + P recombine to form EP and then ES, occurs so rarely that we can ignore it. In this case, EP need not be represented, and we can express the rate of the reaction — known as its velocity, $V$, as

$$V = k_{cat}\,[ES]$$

where [ES] is the concentration of the enzyme–substrate complex, and $k_{cat}$ is the turnover number, a rate constant that has a value equal to the number of substrate molecules processed per enzyme molecule each second.

But how does the value of [ES] relate to the concentrations that we know directly, which are the total concentration of the enzyme, $[E_o]$, and the concentration of the substrate, [S]? When enzyme and substrate are first mixed, the concentration [ES] will rise rapidly from zero to a so-called steady-state level, as illustrated below.




At this steady state, [ES] is nearly constant, so that

$$\text{rate of ES breakdown } k_{-1}\,[ES] + k_{cat}\,[ES] = \text{rate of ES formation } k_1\,[E][S]$$

or, since the concentration of the free enzyme, [E], is equal to $[E_o] - [ES]$,

$$[ES] = \left(\frac{k_1}{k_{-1} + k_{cat}}\right)[E][S] = \left(\frac{k_1}{k_{-1} + k_{cat}}\right)\left([E_o] - [ES]\right)[S]$$

Rearranging, and defining the constant $K_m$ as

$$\frac{k_{-1} + k_{cat}}{k_1}$$

we get

$$[ES] = \frac{[E_o][S]}{K_m + [S]}$$

or, remembering that $V = k_{cat}\,[ES]$, we obtain the famous Michaelis–Menten equation

$$V = \frac{k_{cat}\,[E_o][S]}{K_m + [S]}$$

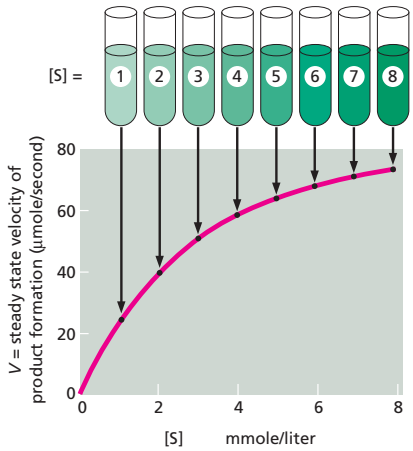
As [S] is increased to higher and higher levels, essentially all of the enzyme will be bound to substrate at steady state; at this point, a maximum rate of reaction, $V_{max}$, will be reached where $V = V_{max} = k_{cat}\,[E_o]$. Thus, it is convenient to rewrite the Michaelis–Menten equation as

$$V = \frac{V_{max}\,[S]}{K_m + [S]}$$

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 44 of 240

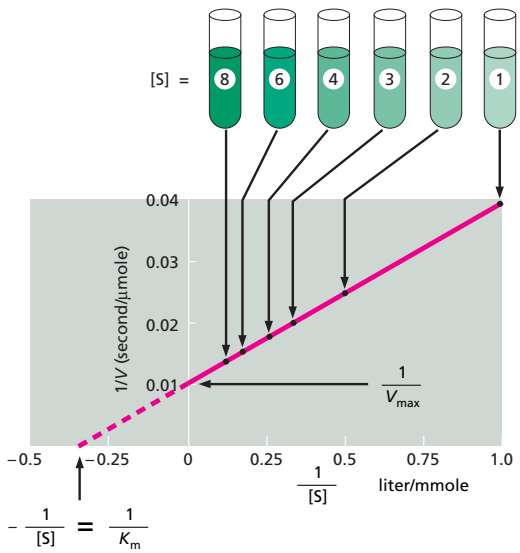
## THE DOUBLE-RECIPROCAL PLOT

A typical plot of $V$ versus [S] for an enzyme that follows Michaelis–Menten kinetics is shown below. From this plot, neither the value of $V_{max}$ nor of $K_m$ is immediately clear.




To obtain $V_{max}$ and $K_m$ from such data, a double-reciprocal plot is often used, in which the Michaelis–Menten equation has merely been rearranged, so that $1/V$ can be plotted versus $1/[S]$.

$$1/V = \left(\frac{K_m}{V_{max}}\right)\left(\frac{1}{[S]}\right) + 1/V_{max}$$




## THE SIGNIFICANCE OF $K_m$, $k_{cat}$, and $k_{cat}/K_m$

As described in the text, $K_m$ is an approximate measure of substrate affinity for the enzyme: it is numerically equal to the concentration of [S] at $V = 0.5\ V_{max}$. In general, a lower value of $K_m$ means tighter substrate binding. In fact, for those cases where $k_{cat}$ is much smaller than $k_{-1}$, the $K_m$ will be equal to $K_d$, the dissociation constant for substrate binding to the enzyme ($K_d = 1/K_a$).

We have seen that $k_{cat}$ is the turnover number for the enzyme. At very low substrate concentrations, where $[S] << K_m$, most of the enzyme is free. Thus we can think of $[E] = [E_o]$, so that the Michaelis–Menten equation becomes $V = k_{cat}/K_m\ [E][S]$. Thus, the ratio $k_{cat}/K_m$ is equivalent to the rate constant for the reaction between free enzyme and free substrate.

A comparison of $k_{cat}/K_m$ for the same enzyme with different substrates, or for two enzymes with their different substrates, is widely used as a measure of enzyme effectiveness.

For simplicity, in this Panel we have discussed enzymes that have only one substrate, such as the lysozyme enzyme described in the text (see p. 164). Most enzymes have two substrates, one of which is often an active carrier molecule—such as NADH or ATP.

A similar, but more complex, analysis is used to determine the kinetics of such enzymes—allowing the order of substrate binding and the presence of covalent intermediates along the pathway to be revealed.

## SOME ENZYMES ARE DIFFUSION LIMITED

The values of $k_{cat}$, $K_m$, and $k_{cat}/K_m$ for some selected enzymes are given below:

| enzyme | substrate | $k_{cat}$ ($sec^{-1}$) | $K_m$ (M) | $k_{at}/K_m$ ($sec^{-1}M^{-1}$) |
|---|---|---|---|---|
| acetylcholinesterase | acetylcholine | $1.4x10^4$ | $9x10^{-5}$ | $1.6x10^8$ |
| catalase | $H_2O_2$ | $4x10^7$ | 1 | $4x10^7$ |
| fumarase | fumarate | $8x10^2$ | $5x10^{-6}$ | $1.6x10^8$ |

Because an enzyme and its substrate must collide before they can react, $k_{cat}/K_m$ has a maximum possible value that is limited by collision rates. If every collision forms an enzyme–substrate complex, one can calculate from diffusion theory that $k_{cat}/K_m$ will be between $10^8$ and $10^9\ sec^{-1}M^{-1}$, in the case where all subsequent steps proceed immediately. Thus, it is claimed that enzymes like acetylcholinesterase and fumarase are "perfect enzymes," each enzyme having evolved to the point where nearly every collision with its substrate converts the substrate to a product.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 45 of 240

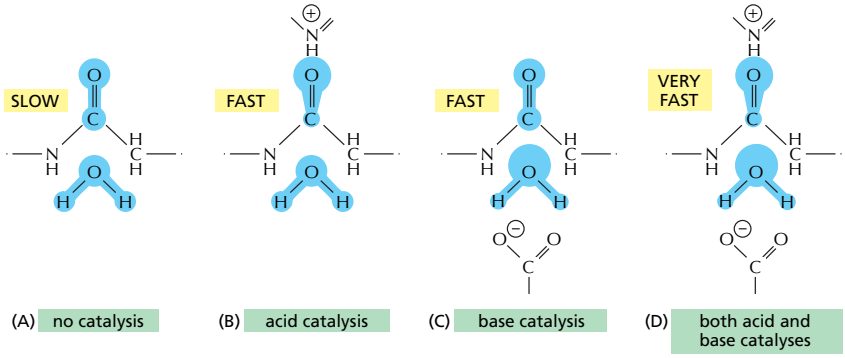



**Figure 3–49 Acid catalysis and base catalysis.** (A) The start of the uncatalyzed reaction shown in Figure 3–47A, with *blue* indicating electron distribution in the water and carbonyl bonds. (B) An acid likes to donate a proton ($H^+$) to other atoms. By pairing with the carbonyl oxygen, an acid causes electrons to move away from the carbonyl carbon, making this atom much more attractive to the electronegative oxygen of an attacking water molecule. (C) A base likes to take up $H^+$. By pairing with a hydrogen of the attacking water molecule, a base causes electrons to move toward the water oxygen, making it a better attacking group for the carbonyl carbon. (D) By having appropriately positioned atoms on its surface, an enzyme can perform both acid catalysis and base catalysis at the same time.

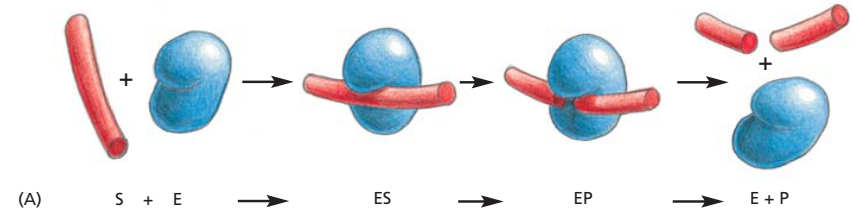
that can be easily isolated in large quantities. For these reasons, it has been intensively studied, and it was the first enzyme to have its structure worked out in atomic detail by x-ray crystallography.

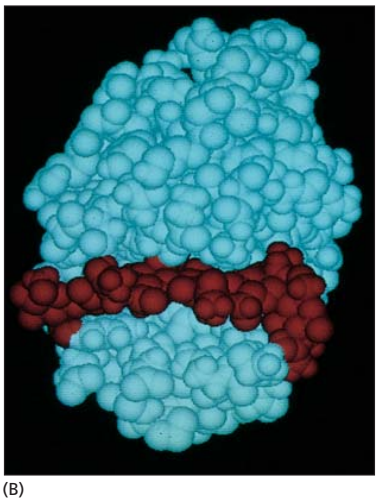
The reaction that lysozyme catalyzes is a hydrolysis: it adds a molecule of water to a single bond between two adjacent sugar groups in the polysaccharide chain, thereby causing the bond to break (see Figure 2–19). The reaction is energetically favorable because the free energy of the severed polysaccharide chain is lower than the free energy of the intact chain. However, the pure polysaccharide can remain for years in water without being hydrolyzed to any detectable degree. This is because there is an energy barrier to the reaction, as discussed in Chapter 2 (see Figure 2–46). A colliding water molecule can break a bond linking two sugars only if the polysaccharide molecule is distorted into a particular shape—the transition state—in which the atoms around the bond have an altered geometry and electron distribution. Because of this distortion, random collisions must supply a very large activation energy for the reaction to take place. In an aqueous solution at room temperature, the energy of collisions almost never exceeds the activation energy. Consequently, hydrolysis occurs extremely slowly, if at all.

This situation changes drastically when the polysaccharide binds to lysozyme. The active site of lysozyme, because its substrate is a polymer, is a long groove that holds six linked sugars at the same time. As soon as the polysaccharide binds to form an enzyme–substrate complex, the enzyme cuts the polysaccharide by adding a water molecule across one of its sugar–sugar bonds. The product chains are then quickly released, freeing the enzyme for further cycles of reaction (**Figure 3–50**).

The chemistry of the binding of lysozyme to its substrate is the same as that for antibody binding to its antigen—the formation of multiple noncovalent

**Figure 3–50 The reaction catalyzed by lysozyme.** (A) The enzyme lysozyme (E) catalyzes the cutting of a polysaccharide chain, which is its substrate (S). The enzyme first binds to the chain to form an enzyme–substrate complex (ES) and then catalyzes the cleavage of a specific covalent bond in the backbone of the polysaccharide, forming an enzyme–product complex (EP) that rapidly dissociates. Release of the severed chain (the products P) leaves the enzyme free to act on another substrate molecule. (B) A space-filling model of the lysozyme molecule bound to a short length of polysaccharide chain before cleavage. (B, courtesy of Richard J. Feldmann.)



(B)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 46 of 240

bonds. However, lysozyme holds its polysaccharide substrate in a particular way, so that it distorts one of the two sugars in the bond to be broken from its normal, most stable conformation. The bond to be broken is also held close to two amino acids with acidic side chains (a glutamic acid and an aspartic acid) within the active site.

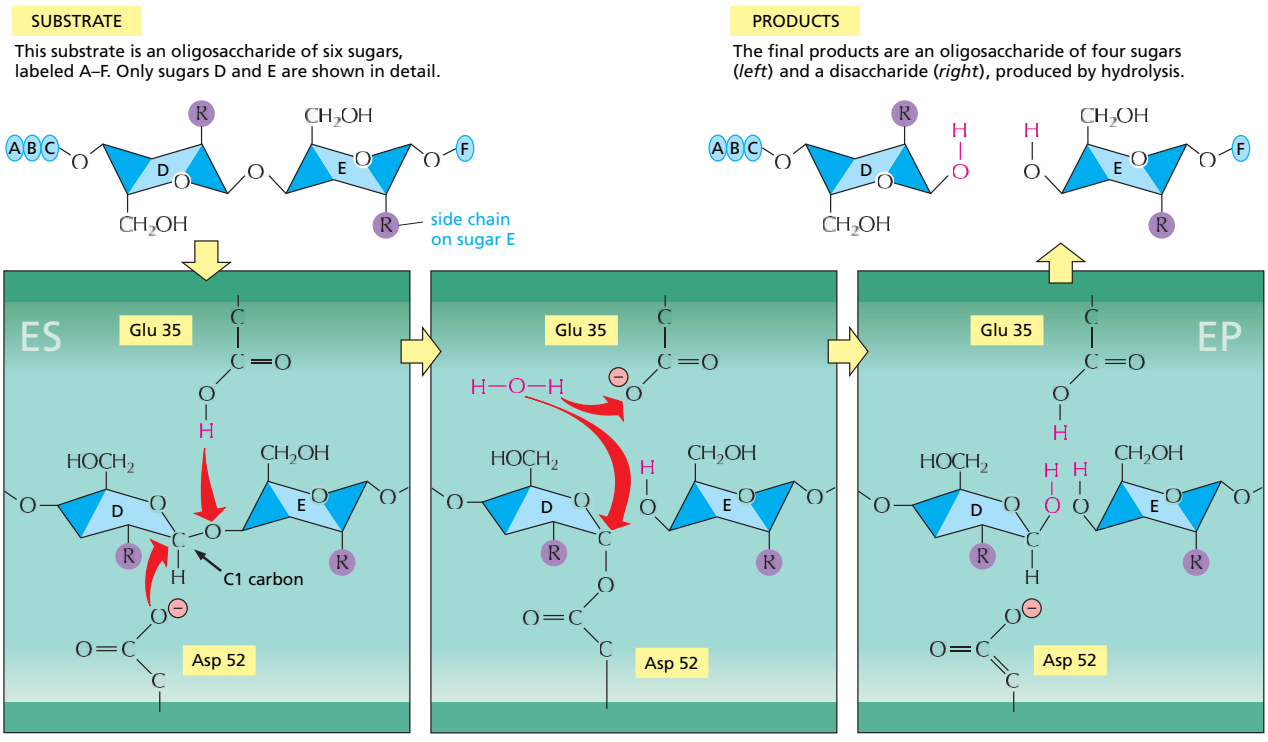
Conditions are thereby created in the microenvironment of the lysozyme active site that greatly reduce the activation energy necessary for the hydrolysis to take place. **Figure 3–51** shows three central steps in this enzymatically catalyzed reaction.

1. The enzyme stresses its bound substrate, so that the shape of one sugar more closely resembles the shape of high-energy transition states formed during the reaction.
2. The negatively charged aspartic acid reacts with the C1 carbon atom on the distorted sugar, and the glutamic acid donates its proton to the oxygen that links this sugar to its neighbor. This breaks the sugar–sugar bond and leaves the aspartic acid side chain covalently linked to the site of bond cleavage.
3. Aided by the negatively charged glutamic acid, a water molecule reacts with the C1 carbon atom, displacing the aspartic acid side chain and completing the process of hydrolysis.

The overall chemical reaction, from the initial binding of the polysaccharide on the surface of the enzyme through the final release of the severed chains, occurs many millions of times faster than it would in the absence of enzyme.

Other enzymes use similar mechanisms to lower activation energies and speed up the reactions they catalyze. In reactions involving two or more reactants, the active site also acts like a template, or mold, that brings the substrates together in the proper orientation for a reaction to occur between them (**Figure**

**Figure 3–51 Events at the active site of lysozyme.** <TGGT> The top left and top right drawings show the free substrate and the free products, respectively, whereas the other three drawings show the sequential events at the enzyme active site. Note the change in the conformation of sugar D in the enzyme–substrate complex; this shape change stabilizes the oxocarbenium ion-like transition states required for formation and hydrolysis of the covalent intermediate shown in the middle panel. It is also possible that a carbonium ion intermediate forms in step 2, as the covalent intermediate shown in the middle panel has been detected only with a synthetic substrate. (See D.J. Vocadlo et al., *Nature* 412:835–838, 2001.)




In the enzyme–substrate complex (ES), the enzyme forces sugar D into a strained conformation, with Glu 35 positioned to serve as an acid that attacks the adjacent sugar–sugar bond by donating a proton ($H^+$) to sugar E, and Asp 52 poised to attack the C1 carbon atom.

The Asp 52 has formed a covalent bond between the enzyme and the C1 carbon atom of sugar D. The Glu 35 then polarizes a water molecule (*red*), so that its oxygen can readily attack the C1 carbon atom and displace Asp 52.

The reaction of the water molecule (*red*) completes the hydrolysis and returns the enzyme to its initial state, forming the final enzyme–product complex (EP).

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 47 of 240

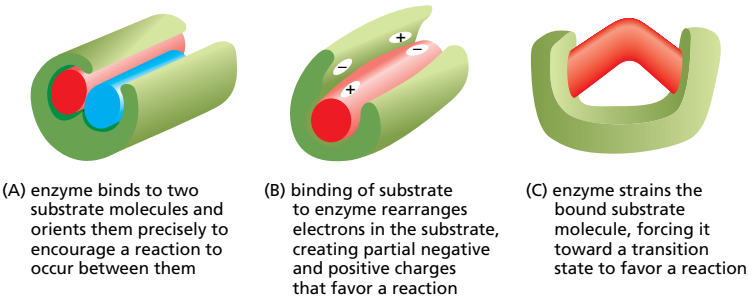



**Figure 3–52 Some general strategies of enzyme catalysis.** (A) Holding substrates together in a precise alignment. (B) Charge stabilization of reaction intermediates. (C) Applying forces that distort bonds in the substrate to increase the rate of a particular reaction.

3–52A). As we saw for lysozyme, the active site of an enzyme contains precisely positioned atoms that speed up a reaction by using charged groups to alter the distribution of electrons in the substrates (Figure 3–52B). In addition, when a substrate binds to an enzyme, bonds in the substrate often bend, changing the substrate shape. These changes, along with mechanical forces, drive a substrate toward a particular transition state (Figure 3–52C). Finally, like lysozyme, many enzymes participate intimately in the reaction by briefly forming a covalent bond between the substrate and a side chain of the enzyme. Subsequent steps in the reaction restore the side chain to its original state, so that the enzyme remains unchanged after the reaction (see also Figure 2–72).

## Tightly Bound Small Molecules Add Extra Functions to Proteins

Although we have emphasized the versatility of proteins as chains of amino acids that perform different functions, there are many instances in which the amino acids by themselves are not enough. Just as humans employ tools to enhance and extend the capabilities of their hands, proteins often use small nonprotein molecules to perform functions that would be difficult or impossible to do with amino acids alone. Thus, the signal receptor protein *rhodopsin*, which is made by the photoreceptor cells in the retina, detects light by means of a small molecule, *retinal*, embedded in the protein (**Figure 3–53**A). Retinal changes its shape when it absorbs a photon of light, and this change causes the protein to trigger a cascade of enzymatic reactions that eventually lead to an electrical signal being carried to the brain.

Another example of a protein that contains a nonprotein portion is hemoglobin (see Figure 3–22). A molecule of hemoglobin carries four *heme* groups, ring-shaped molecules each with a single central iron atom (Figure 3–53B). Heme gives hemoglobin (and blood) its red color. By binding reversibly to oxygen gas through its iron atom, heme enables hemoglobin to pick up oxygen in the lungs and release it in the tissues.

Sometimes these small molecules are attached covalently and permanently to their protein, thereby becoming an integral part of the protein molecule itself. We shall see in Chapter 10 that proteins are often anchored to cell membranes through covalently attached lipid molecules. And membrane proteins exposed

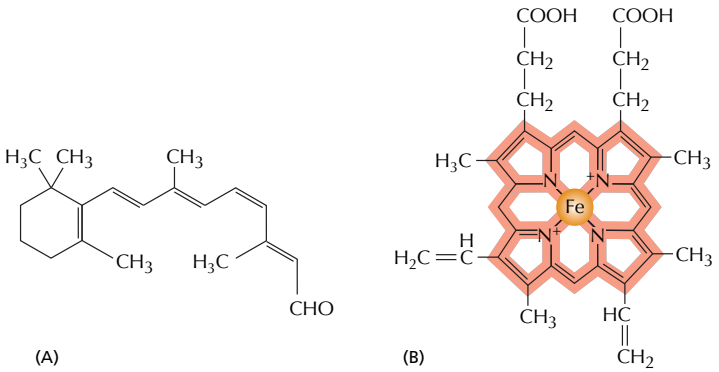



**Figure 3–53 Retinal and heme.** (A) The structure of retinal, the light-sensitive molecule attached to rhodopsin in the eye. (B) The structure of a heme group. The carbon-containing heme ring is *red* and the iron atom at its center is *orange*. A heme group is tightly bound to each of the four polypeptide chains in hemoglobin, the oxygen-carrying protein whose structure is shown in Figure 3–22.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 48 of 240

**Table 3–2 Many Vitamins Provide Critical Coenzymes for Human Cells**

| VITAMIN | COENZYME | ENZYME-CATALYZED REACTIONS REQUIRING THESE COENZYMES |
|---|---|---|
| Thiamine (vitamin $B_1$) | thiamine pyrophosphate | activation and transfer of aldehydes |
| Riboflavin (vitamin $B_2$) | FADH | oxidation–reduction |
| Niacin | NADH, NADPH | oxidation–reduction |
| Pantothenic acid | coenzyme A | acyl group activation and transfer |
| Pyridoxine | pyridoxal phosphate | amino acid activation; also glycogen phosphorylase |
| Biotin | biotin | $CO_2$ activation and transfer |
| Lipoic acid | lipoamide | acyl group activation; oxidation–reduction |
| Folic acid | tetrahydrofolate | activation and transfer of single carbon groups |
| Vitamin $B_{12}$ | cobalamin coenzymes | isomerization and methyl group transfers |

on the surface of the cell, as well as proteins secreted outside the cell, are often modified by the covalent addition of sugars and oligosaccharides.

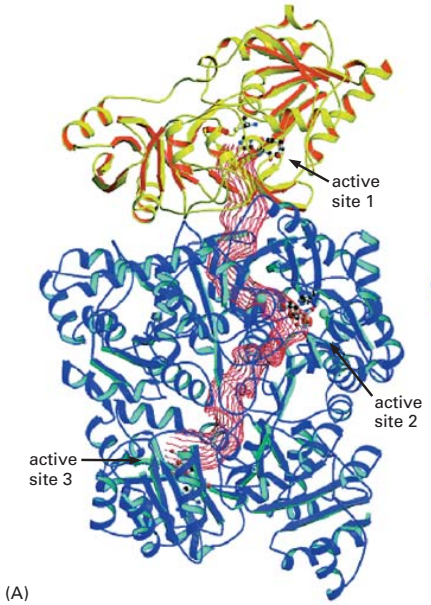
Enzymes frequently have a small molecule or metal atom tightly associated with their active site that assists with their catalytic function. *Carboxypeptidase*, for example, an enzyme that cuts polypeptide chains, carries a tightly bound zinc ion in its active site. During the cleavage of a peptide bond by carboxypeptidase, the zinc ion forms a transient bond with one of the substrate atoms, thereby assisting the hydrolysis reaction. In other enzymes, a small organic molecule serves a similar purpose. Such organic molecules are often referred to as **coenzymes**. An example is *biotin*, which is found in enzymes that transfer a carboxylate group ($-COO^-$) from one molecule to another (see Figure 2–63). Biotin participates in these reactions by forming a transient covalent bond to the $-COO^-$ group to be transferred, being better suited to this function than any of the amino acids used to make proteins. Because it cannot be synthesized by humans, and must therefore be supplied in small quantities in our diet, biotin is a *vitamin*. Many other coenzymes are produced from vitamins (**Table 3–2**). Vitamins are also needed to make other types of small molecules that are essential components of our proteins; vitamin A, for example, is needed in the diet to make retinal, the light-sensitive part of rhodopsin.

## Molecular Tunnels Channel Substrates in Enzymes with Multiple Catalytic Sites

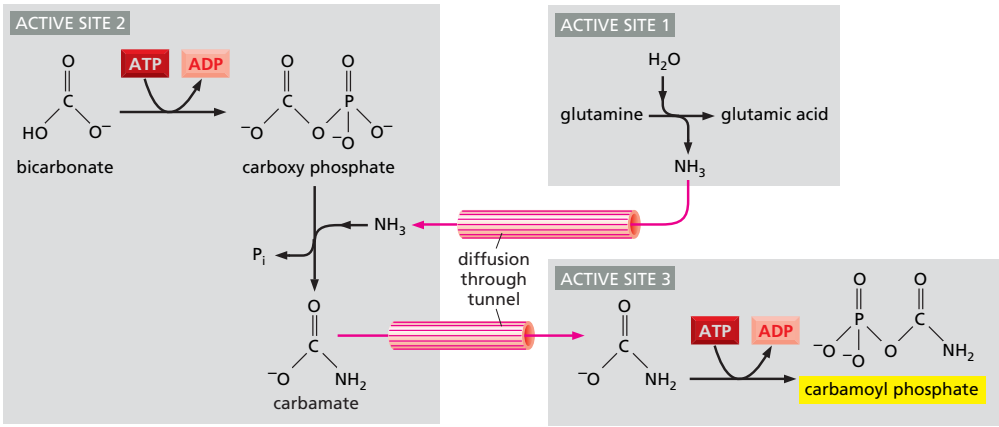
Some of the chemical reactions catalyzed by enzymes in cells produce intermediates that are either very unstable or that could readily diffuse out of the cell through the plasma membrane if released into the cytosol. To preserve these intermediates, enzymes have evolved *molecular tunnels* that connect two or more active sites, allowing the intermediate to be rapidly processed to a final product—without ever leaving the enzyme.

Consider, for example, the enzyme carbamoyl phosphate synthetase, which uses ammonia derived from glutamine plus two molecules of ATP to convert bicarbonate ($HCO_3^-$) to carbamoyl phosphate—an important intermediate in several metabolic pathways (**Figure 3–54**). This enzyme contains three widely separated active sites that are connected to each other by a tunnel. The reaction starts at active site 2, located in the middle of the tunnel, where ATP is used to phosphorylate (add a phosphate group to) bicarbonate, forming carboxy phosphate. This event triggers the hydrolysis of glutamine to glutamic acid at active site 1, releasing ammonia into the tunnel. The ammonia immediately diffuses through the first half of the tunnel to active site 2, where it reacts with the carboxyphosphate to form carbamate. This unstable intermediate then diffuses through the second half of the tunnel to active site 3, where it is phosphorylated by ATP to the final product, carbamoyl phosphate.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 49 of 240




(A)




(B)

**Figure 3–54 The tunneling of reaction intermediates in the enzyme carbamoyl phosphate synthetase.** (A) Diagram of the structure of the enzyme, in which a *red* ribbon has been used to outline the tunnel on the inside of the protein connecting its three active sites. The small and large subunits of this dimeric enzyme are color coded *yellow* and *blue*, respectively. (B) The path of the reaction. As indicated, active site 1 produces ammonia, which diffuses through the tunnel to active site 2, where it combines with carboxy phosphate to form carbamate. This highly unstable intermediate then diffuses through the tunnel to active site 3, where it is phosphorylated by ATP to produce the final product, carbamoyl phosphate. (A, modified from F.M. Raushel, J.B. Thoden, and H.M. Holden, *Acc. Chem. Res.* 36:539–548, 2003. With permission from American Chemical Society.)

Several other well characterized enzymes contain similar molecular tunnels. Ammonia, a readily diffusable intermediate that might otherwise be lost from the cell, is the substrate most frequently channeled in the examples thus far known.

## Multienzyme Complexes Help to Increase the Rate of Cell Metabolism

The efficiency of enzymes in accelerating chemical reactions is crucial to the maintenance of life. Cells, in effect, must race against the unavoidable processes of decay, which—if left unattended—cause macromolecules to run downhill toward greater and greater disorder. If the rates of desirable reactions were not greater than the rates of competing side reactions, a cell would soon die. We can get some idea of the rate at which cell metabolism proceeds by measuring the rate of ATP utilization. A typical mammalian cell "turns over" (i.e., hydrolyzes and restores by phosphorylation) its entire ATP pool once every 1 or 2 minutes. For each cell, this turnover represents the utilization of roughly $10^7$ molecules of ATP per second (or, for the human body, about 1 gram of ATP every minute).

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 50 of 240

The rates of reactions in cells are rapid because enzyme catalysis is so effective. Many important enzymes have become so efficient that there is no possibility of further useful improvement. The factor that limits the reaction rate is no longer the enzyme's intrinsic speed of action; rather, it is the frequency with which the enzyme collides with its substrate. Such a reaction is said to be *diffusion-limited* (see Panel 3–3, p. 162–163).

If an enzyme-catalyzed reaction is diffusion-limited, its rate depends on the concentration of both the enzyme and its substrate. If a sequence of reactions is to occur extremely rapidly, each metabolic intermediate and enzyme involved must be present in high concentration. However, given the enormous number of different reactions performed by a cell, there are limits to the concentrations that can be achieved. In fact, most metabolites are present in micromolar ($10^{-6}$ M) concentrations, and most enzyme concentrations are much lower. How is it possible, therefore, to maintain very fast metabolic rates?

The answer lies in the spatial organization of cell components. The cell can increase reaction rates without raising substrate concentrations by bringing the various enzymes involved in a reaction sequence together to form a large protein assembly known as a *multienzyme complex* (**Figure 3–55**). Because this allows the product of enzyme A to be passed directly to enzyme B, and so on, diffusion rates need not be limiting, even when the concentrations of the substrates in the cell as a whole are very low. It is perhaps not surprising, therefore, that such enzyme complexes are very common, and they are involved in nearly all aspects of metabolism—including the central genetic processes of DNA, RNA, and protein synthesis. In fact, few enzymes in eucaryotic cells diffuse freely in solution; instead, most seem to have evolved binding sites that concentrate them with other proteins of related function in particular regions of the cell, thereby increasing the rate and efficiency of the reactions that they catalyze.

Eucaryotic cells have yet another way of increasing the rate of metabolic reactions: using their intracellular membrane systems. These membranes can segregate particular substrates and the enzymes that act on them into the same membrane-enclosed compartment, such as the endoplasmic reticulum or the cell nucleus. If, for example, a compartment occupies a total of 10% of the volume of the cell, the concentration of reactants in that compartment may be increased by 10 times compared with a cell with the same number of enzymes and substrate molecules, but no compartmentalization. Reactions limited by the speed of diffusion can thereby be speeded up by a factor of 10.

## The Cell Regulates the Catalytic Activities of its Enzymes

A living cell contains thousands of enzymes, many of which operate at the same time and in the same small volume of the cytosol. By their catalytic action, these enzymes generate a complex web of metabolic pathways, each composed of chains of chemical reactions in which the product of one enzyme becomes the substrate of the next. In this maze of pathways, there are many branch points (nodes) where different enzymes compete for the same substrate. The system is so complex (see Figure 2–88) that elaborate controls are required to regulate when and how rapidly each reaction occurs.

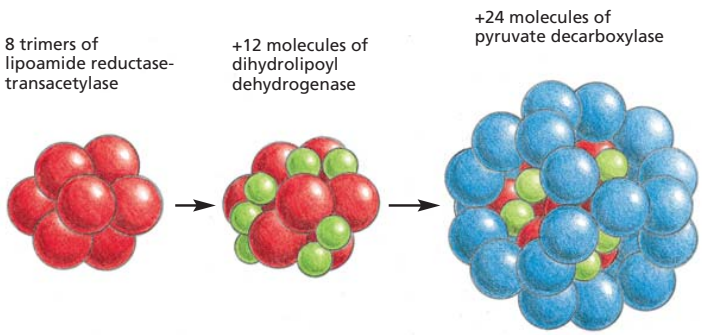



**Figure 3–55 The structure of pyruvate dehydrogenase.** This enzyme complex catalyzes the conversion of pyruvate to acetyl CoA, as part of the pathway that oxidizes sugars to $CO_2$ and $H_2O$ (see Figure 2–79). It is an example of a large multienzyme complex in which reaction intermediates are passed directly from one enzyme to another.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 51 of 240

Regulation occurs at many levels. At one level, the cell controls how many molecules of each enzyme it makes by regulating the expression of the gene that encodes that enzyme (discussed in Chapter 7). The cell also controls enzymatic activities by confining sets of enzymes to particular subcellular compartments, enclosed by distinct membranes (discussed in Chapters 12 and 14). As will be discussed later in this chapter, enzymes are frequently covalently modified to control their activity. The rate of protein destruction by targeted proteolysis represents yet another important regulatory mechanism (see p. 395). But the most general process that adjusts reaction rates operates through a direct, reversible change in the activity of an enzyme in response to the specific small molecules that it encounters.

The most common type of control occurs when a molecule other than one of the substrates binds to an enzyme at a special regulatory site outside the active site, thereby altering the rate at which the enzyme converts its substrates to products. For example, in **feedback inhibition** a product produced late in a reaction pathway inhibits an enzyme that acts earlier in the pathway. Thus, whenever large quantities of the final product begin to accumulate, this product binds to the enzyme and slows down its catalytic action, thereby limiting the further entry of substrates into that reaction pathway (**Figure 3–56**). Where pathways branch or intersect, there are usually multiple points of control by different final products, each of which works to regulate its own synthesis (**Figure 3–57**). Feedback inhibition can work almost instantaneously, and it is rapidly reversed when the level of the product falls.

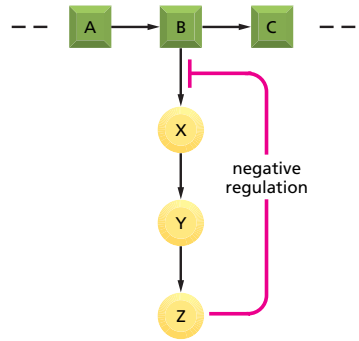



**Figure 3–56 Feedback inhibition of a single biosynthetic pathway.** The end-product Z inhibits the first enzyme that is unique to its synthesis and thereby controls its own level in the cell. This is an example of negative regulation.

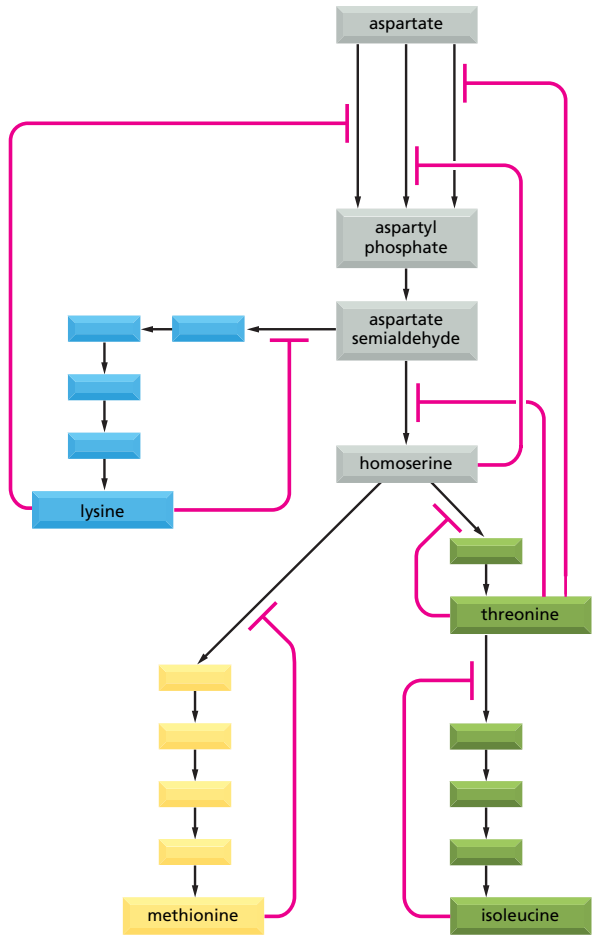



**Figure 3–57 Multiple feedback inhibition.** In this example, which shows the biosynthetic pathways for four different amino acids in bacteria, the *red arrows* indicate positions at which products feed back to inhibit enzymes. Each amino acid controls the first enzyme specific to its own synthesis, thereby controlling its own levels and avoiding a wasteful, or even dangerous, buildup of intermediates. The products can also separately inhibit the initial set of reactions common to all the syntheses; in this case, three different enzymes catalyze the initial reaction, each inhibited by a different product.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 52 of 240

Feedback inhibition is *negative regulation:* it prevents an enzyme from acting. Enzymes can also be subject to *positive regulation*, in which a regulatory molecule stimulates the enzyme's activity rather than shutting the enzyme down. Positive regulation occurs when a product in one branch of the metabolic network stimulates the activity of an enzyme in another pathway. As one example, the accumulation of ADP activates several enzymes involved in the oxidation of sugar molecules, thereby stimulating the cell to convert more ADP to ATP.

## Allosteric Enzymes Have Two or More Binding Sites That Interact

A striking feature of both positive and negative feedback regulation is that the regulatory molecule often has a shape totally different from the shape of the substrate of the enzyme. This is why the effect on a protein is termed *allostery* (from the Greek words *allos*, meaning "other," and *stereos*, meaning "solid" or "three-dimensional"). As biologists learned more about feedback regulation, they recognized that the enzymes involved must have at least two different binding sites on their surface—an **active site** that recognizes the substrates, and a **regulatory site** that recognizes a regulatory molecule. These two sites must somehow communicate so that the catalytic events at the active site can be influenced by the binding of the regulatory molecule at its separate site on the protein's surface.

The interaction between separated sites on a protein molecule is now known to depend on a *conformational change* in the protein: binding at one of the sites causes a shift from one folded shape to a slightly different folded shape. During feedback inhibition, for example, the binding of an inhibitor at one site on the protein causes the protein to shift to a conformation that incapacitates its active site, located elsewhere in the protein.

It is thought that most protein molecules are allosteric. They can adopt two or more slightly different conformations, and a shift from one to another caused by the binding of a ligand can alter their activity. This is true not only for enzymes but also for many other proteins, including receptors, structural proteins, and motor proteins. In all instances of allosteric regulation, each conformation of the protein has somewhat different surface contours, and the protein's binding sites for ligands are altered when the protein changes shape. Moreover as we discuss next, each ligand will stabilize the conformation that it binds to most strongly, and thus—at high enough concentrations—will tend to "switch" the protein toward the conformation that the ligand prefers.

## Two Ligands Whose Binding Sites Are Coupled Must Reciprocally Affect Each Other's Binding

The effects of ligand binding on a protein follow from a fundamental chemical principle known as **linkage**. Suppose, for example, that a protein that binds glucose also binds another molecule, X, at a distant site on the protein's surface. If the binding site for X changes shape as part of the conformational change induced by glucose binding, the binding sites for X and for glucose are said to be *coupled*. Whenever two ligands prefer to bind to the *same* conformation of an allosteric protein, it follows from basic thermodynamic principles that each ligand must increase the affinity of the protein for the other. Thus, if the shift of the protein in **Figure 3–58** to the closed conformation that binds glucose best also causes the binding site for X to fit X better, then the protein will bind glucose more tightly when X is present than when X is absent.

Conversely, linkage operates in a negative way if two ligands prefer to bind to *different* conformations of the same protein. In this case, the binding of the first ligand discourages the binding of the second ligand. Thus, if a shape change caused by glucose binding decreases the affinity of a protein for molecule X, the binding of X must also decrease the protein's affinity for glucose (**Figure 3–59**). The linkage relationship is quantitatively reciprocal, so that, for example, if glucose has a very large effect on the binding of X, X has a very large effect on the binding of glucose.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 53 of 240

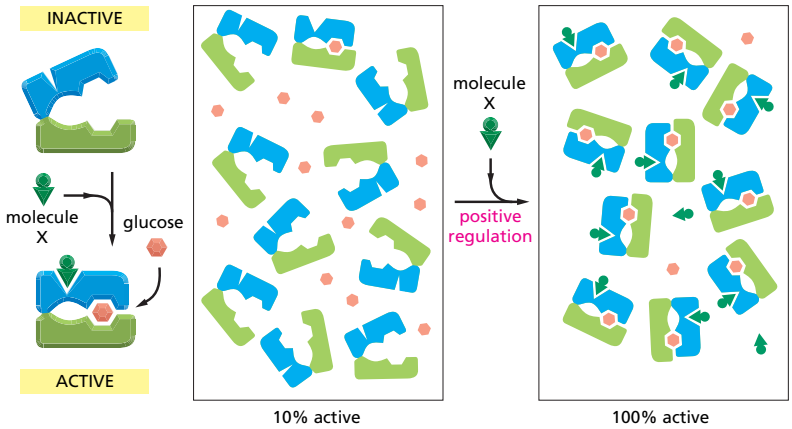



**Figure 3–58 Positive regulation caused by conformational coupling between two distant binding sites.** In this example, both glucose and molecule X bind best to the *closed* conformation of a protein with two domains. Because both glucose and molecule X drive the protein toward its closed conformation, each ligand helps the other to bind. Glucose and molecule X are therefore said to bind *cooperatively* to the protein.

The relationships shown in Figures 3–58 and 3–59 apply to all proteins, and they underlie all of cell biology. They seem so obvious in retrospect that we now take it for granted. But the discovery of linkage in studies of a few enzymes in the 1950s, followed by an extensive analysis of allosteric mechanisms in proteins in the early 1960s, had a revolutionary effect on our understanding of biology. Since molecule X in these examples binds at a site on the enzyme that is distinct from the site where catalysis occurs, it need have no chemical relationship to glucose or to any other ligand that binds at the active site. Moreover, as we have just seen, for enzymes that are regulated in this way, molecule X can either turn the enzyme on (positive regulation) or turn it off (negative regulation). By such a mechanism, **allosteric proteins** serve as general switches that, in principle, allow one molecule in a cell to affect the fate of any other.

## Symmetric Protein Assemblies Produce Cooperative Allosteric Transitions

A single-subunit enzyme that is regulated by negative feedback can at most decrease from 90% to about 10% activity in response to a 100-fold increase in the concentration of an inhibitory ligand that it binds (**Figure 3–60**, *red line).* Responses of this type are apparently not sharp enough for optimal cell regulation, and most enzymes that are turned on or off by ligand binding consist of symmetric assemblies of identical subunits. With this arrangement, the binding of a molecule of ligand to a single site on one subunit can promote an allosteric change in the entire assembly that helps the neighboring subunits bind the same ligand. As a result, a *cooperative allosteric transition* occurs (Figure 3–60, *blue line),* allowing a relatively small change in ligand concentration in the cell to switch the whole assembly from an almost fully active to an almost fully inactive conformation (or vice versa).

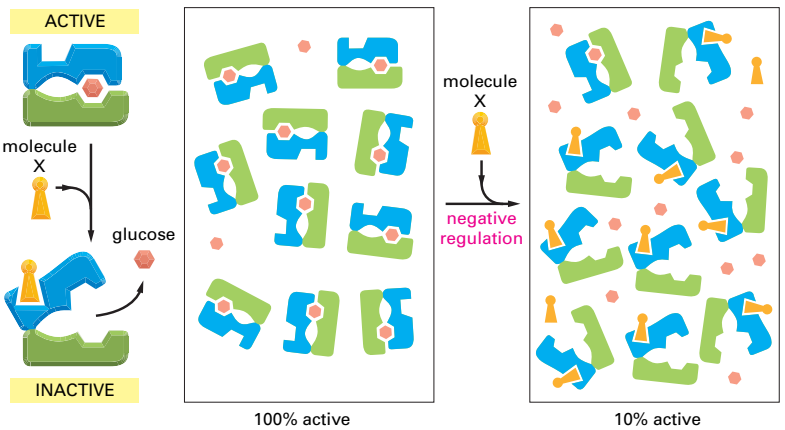



**Figure 3–59 Negative regulation caused by conformational coupling between two distant binding sites.** The scheme here resembles that in the previous figure, but here molecule X prefers the *open* conformation, while glucose prefers the *closed* conformation. Because glucose and molecule X drive the protein toward opposite conformations (closed and open, respectively), the presence of either ligand interferes with the binding of the other.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 54 of 240

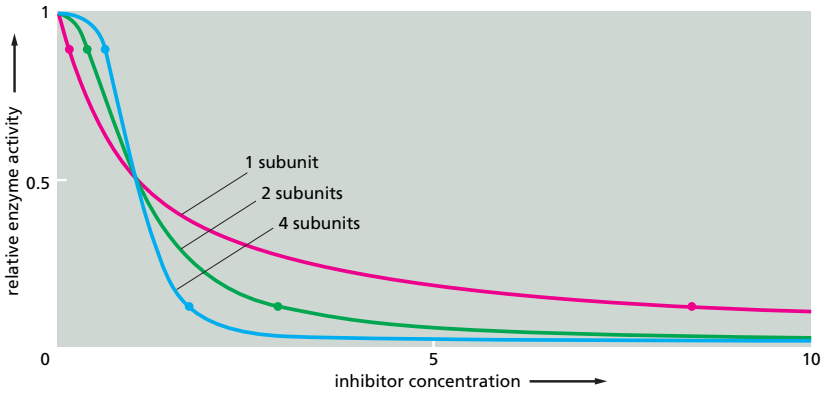



**Figure 3–60 Enzyme activity versus the concentration of inhibitory ligand for single-subunit and multisubunit allosteric enzymes.** For an enzyme with a single subunit *(red line),* a drop from 90% enzyme activity to 10% activity (indicated by the two dots on the curve) requires a 100-fold increase in the concentration of inhibitor. The enzyme activity is calculated from the simple equilibrium relationship $K = [IP]/[I][P]$, where P is active protein, I is inhibitor, and IP is the inactive protein bound to inhibitor. An identical curve applies to any simple binding interaction between two molecules, A and B. In contrast, a multisubunit allosteric enzyme can respond in a switchlike manner to a change in ligand concentration: the steep response is caused by a cooperative binding of the ligand molecules, as explained in Figure 3–61. Here, the *green line* represents the idealized result expected for the cooperative binding of two inhibitory ligand molecules to an allosteric enzyme with two subunits, and the *blue line* shows the idealized response of an enzyme with four subunits. As indicated by the two dots on each of these curves, the more complex enzymes drop from 90% to 10% activity over a much narrower range of inhibitor concentration than does the enzyme composed of a single subunit.

The principles involved in a cooperative "all-or-none" transition are the same for all proteins, whether or not they are enzymes. But they are perhaps easiest to visualize for an enzyme that forms a symmetric dimer. In the example shown in **Figure 3–61**, the first molecule of an inhibitory ligand binds with great difficulty since its binding disrupts an energetically favorable interaction between the two identical monomers in the dimer. A second molecule of inhibitory ligand now binds more easily, however, because its binding restores the energetically favorable monomer–monomer contacts of a symmetric dimer (this also completely inactivates the enzyme).

As an alternative to this *induced fit* model for a cooperative allosteric transition, we can view such a symmetrical enzyme as having only two possible conformations, corresponding to the "enzyme on" and "enzyme off" structures in Figure 3–61. In this view, ligand binding perturbs an all-or-none equilibrium between these two states, thereby changing the proportion of active molecules. Both models represent true and useful concepts; it is the second model that we shall describe next.

## The Allosteric Transition in Aspartate Transcarbamoylase Is Understood in Atomic Detail

One enzyme used in the early studies of allosteric regulation was aspartate transcarbamoylase from *E. coli*. It catalyzes the important reaction that begins the synthesis of the pyrimidine ring of C, U, and T nucleotides: carbamoyl phosphate + aspartate → *N*-carbamoylaspartate. One of the final products of this pathway, cytosine triphosphate (CTP), binds to the enzyme to turn it off whenever CTP is plentiful.

Aspartate transcarbamoylase is a large complex of six regulatory and six catalytic subunits. The catalytic subunits form two trimers, each arranged in the shape of an equilateral triangle; the two trimers face each other and are held

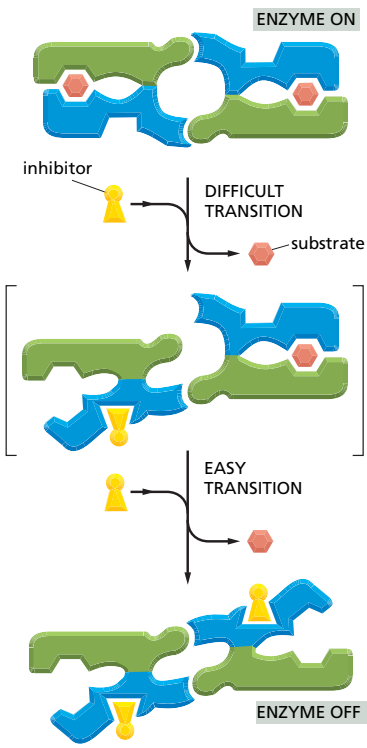



**Figure 3–61 A cooperative allosteric transition in an enzyme composed of two identical subunits.** This diagram illustrates how the conformation of one subunit can influence that of its neighbor. The binding of a single molecule of an inhibitory ligand *(yellow)* to one subunit of the enzyme occurs with difficulty because it changes the conformation of this subunit and thereby disrupts the symmetry of the enzyme. Once this conformational change has occurred, however, the energy gained by restoring the symmetric pairing interaction between the two subunits makes it especially easy for the second subunit to bind the inhibitory ligand and undergo the same conformational change. Because the binding of the first molecule of ligand increases the affinity with which the other subunit binds the same ligand, the response of the enzyme to changes in the concentration of the ligand is much steeper than the response of an enzyme with only one subunit (see Figure 3–60).

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 55 of 240

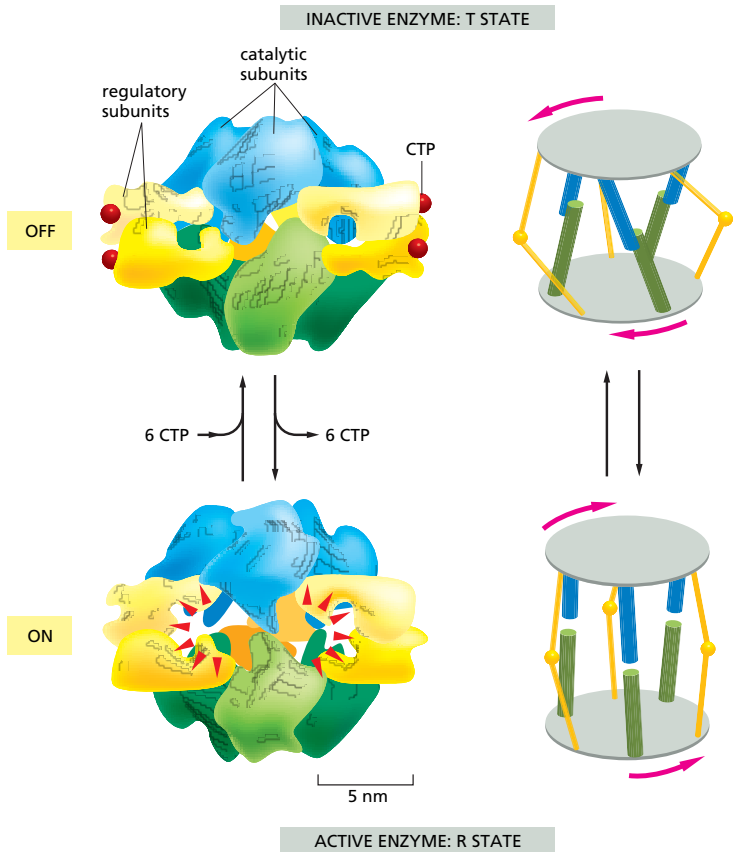



**Figure 3–62 The transition between R and T states in the enzyme aspartate transcarbamoylase.** <CTAA> The enzyme consists of a complex of six catalytic subunits and six regulatory subunits, and the structures of its inactive (T state) and active (R state) forms have been determined by x-ray crystallography. The enzyme is turned off by feedback inhibition when CTP concentrations rise. Each regulatory subunit can bind one molecule of CTP, which is one of the final products in the pathway. By means of this negative feedback regulation, the pathway is prevented from producing more CTP than the cell needs. (Based on K.L. Krause, K.W. Volz and W.N. Lipscomb, *Proc. Natl Acad. Sci. U.S.A.* 82:1643–1647, 1985. With permission from National Academy of Sciences.)

together by three regulatory dimers that form a bridge between them. The entire molecule is poised to undergo a concerted, all-or-none, allosteric transition between two conformations, designated as T (tense) and R (relaxed) states (**Figure 3–62**).

The binding of substrates (carbamoyl phosphate and aspartate) to the catalytic trimers drives aspartate transcarbamoylase into its catalytically active R state, from which the regulatory CTP molecules dissociate. By contrast, the binding of CTP to the regulatory dimers converts the enzyme to the inactive T state, from which the substrates dissociate. This tug-of-war between CTP and substrates is identical in principle to that described previously in Figure 3–59 for a simpler allosteric protein. But because the tug-of-war occurs in a symmetric molecule with multiple binding sites, the enzyme undergoes a cooperative allosteric transition that will turn it on suddenly as substrates accumulate (forming the R state) or shut it off rapidly when CTP accumulates (forming the T state).

A combination of biochemistry and x-ray crystallography has revealed many fascinating details of this allosteric transition. Each regulatory subunit has two domains, and the binding of CTP causes the two domains to move relative to each other, so that they function like a lever that rotates the two catalytic trimers and pulls them closer together into the T state (see Figure 3–62). When this occurs, hydrogen bonds form between opposing catalytic subunits. This helps widen the cleft that forms the active site within each catalytic subunit, thereby disrupting the binding sites for the substrates (**Figure 3–63**). Adding large amounts of substrate has the opposite effect, favoring the R state by binding in the cleft of each catalytic subunit and opposing the above conformational change. Conformations that are intermediate between R and T are unstable, so that the enzyme mostly clicks back and forth between its R and T forms, producing a mixture of these two species in proportions that depend on the relative concentrations of CTP and substrates.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 56 of 240

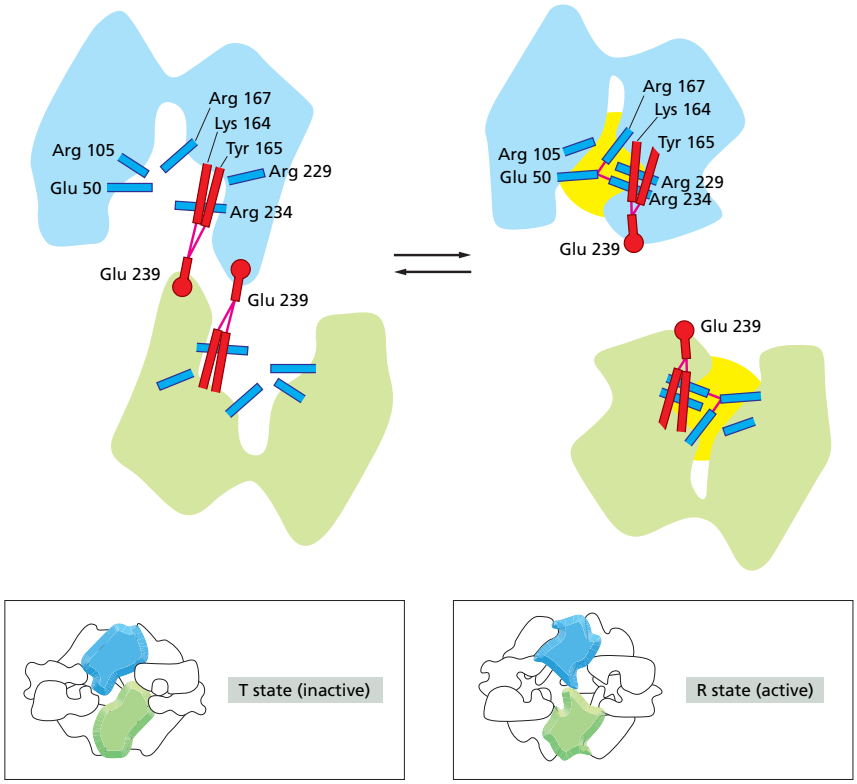



**Figure 3–63 Part of the on–off switch in the catalytic subunits of aspartate transcarbamoylase.** Changes in the indicated hydrogen-bonding interactions are partly responsible for switching this enzyme's active site between active *(yellow)* and inactive conformations. Hydrogen bonds are indicated by *thin red lines*. The amino acids involved in the subunit–subunit interaction in the T state are shown in *red*, while those that form the active site of the enzyme in the R state are shown in *blue*. The large drawings show the catalytic site in the interior of the enzyme; the boxed sketches show the same subunits viewed from the enzyme's external surface. (Adapted from E.R. Kantrowitz and W.N. Lipscomb, *Trends Biochem. Sci.* 15:53–59, 1990. With permission from Elsevier.)

## Many Changes in Proteins Are Driven by Protein Phosphorylation

Proteins are regulated by more than the reversible binding of other molecules. A second method that eucaryotic cells use to regulate a protein's function is the covalent addition of a smaller molecule to one or more of its amino acid side chains. The most common such regulatory modification in higher eucaryotes is the addition of a phosphate group. We shall therefore use protein phosphorylation to illustrate some of the general principles involved in the control of protein function through the modification of amino acid side chains.

A phosphorylation event can affect the protein that is modified in two important ways. First, because each phosphate group carries two negative charges, the enzyme-catalyzed addition of a phosphate group to a protein can cause a major conformational change in the protein by, for example, attracting a cluster of positively charged amino acid side chains. This can, in turn, affect the binding of ligands elsewhere on the protein surface, dramatically changing the protein's activity. When a second enzyme removes the phosphate group, the protein returns to its original conformation and restores its initial activity.

Second, an attached phosphate group can form part of a structure that the binding sites of other proteins recognize. As previously discussed, certain protein domains, sometimes referred to as modules, appear very frequently as parts of larger proteins. One such module is the SH2 domain, described earlier, which binds to a short peptide sequence containing a phosphorylated tyrosine side chain (see Figure 3–39B). More than ten other common domains provide binding sites for attaching their protein to phosphorylated peptides in other protein molecules, each recognizing a phosphorylated amino acid side chain in a different protein context. As a result, protein phosphorylation and dephosphorylation very often drive the regulated assembly and disassembly of protein complexes (see Figure 15–22).

Reversible protein phosphorylation controls the activity, structure, and cellular localization of both enzymes and many other types of proteins in

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 57 of 240

eucaryotic cells. In fact, this regulation is so extensive that more than one-third of the 10,000 or so proteins in a typical mammalian cell are thought to be phosphorylated at any given time—many with more than one phosphate. As might be expected, the addition and removal of phosphate groups from specific proteins often occur in response to signals that specify some change in a cell's state. For example, the complicated series of events that takes place as a eucaryotic cell divides is largely timed in this way (discussed in Chapter 17), and many of the signals mediating cell–cell interactions are relayed from the plasma membrane to the nucleus by a cascade of protein phosphorylation events (discussed in Chapter 15).

## A Eucaryotic Cell Contains a Large Collection of Protein Kinases and Protein Phosphatases

Protein phosphorylation involves the enzyme-catalyzed transfer of the terminal phosphate group of an ATP molecule to the hydroxyl group on a serine, threonine, or tyrosine side chain of the protein (**Figure 3–64**). A **protein kinase** catalyzes this reaction, and the reaction is essentially unidirectional because of the large amount of free energy released when the phosphate–phosphate bond in ATP is broken to produce ADP (discussed in Chapter 2). A **protein phosphatase** catalyzes the reverse reaction of phosphate removal, or *dephosphorylation*. Cells contain hundreds of different protein kinases, each responsible for phosphorylating a different protein or set of proteins. There are also many different protein phosphatases; some are highly specific and remove phosphate groups from only one or a few proteins, whereas others act on a broad range of proteins and are targeted to specific substrates by regulatory subunits. The state of phosphorylation of a protein at any moment, and thus its activity, depends on the relative activities of the protein kinases and phosphatases that modify it.

The protein kinases that phosphorylate proteins in eucaryotic cells belong to a very large family of enzymes, which share a catalytic (kinase) sequence of about 290 amino acids. The various family members contain different amino acid sequences on either end of the kinase sequence (for example, see Figure 3–10), and often have short amino acid sequences inserted into loops within it *(red arrowheads* in **Figure 3–65**). Some of these additional amino acid sequences enable each kinase to recognize the specific set of proteins it phosphorylates, or to bind to structures that localize it in specific regions of the cell. Other parts of the protein regulate the activity of each kinase, so it can be turned on and off in response to different specific signals, as described below.

By comparing the number of amino acid sequence differences between the various members of a protein family, we can construct an "evolutionary tree" that is thought to reflect the pattern of gene duplication and divergence that gave rise to the family. **Figure 3–66** shows an evolutionary tree of protein kinases. Kinases with related functions are often located on nearby branches of the tree: the protein kinases involved in cell signaling that phosphorylate tyrosine side chains, for example, are all clustered in the top left corner of the tree. The other kinases shown phosphorylate either a serine or a threonine side chain, and many are organized into clusters that seem to reflect their function—in transmembrane signal transduction, intracellular signal amplification, cell-cycle control, and so on.

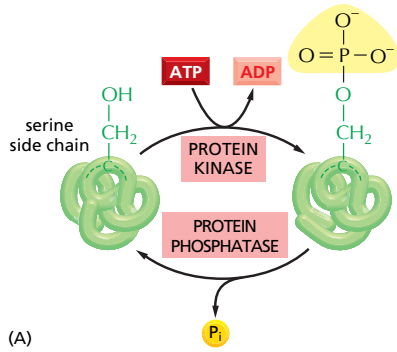



**Figure 3–64 Protein phosphorylation.** Many thousands of proteins in a typical eucaryotic cell are modified by the covalent addition of a phosphate group. (A) The general reaction, shown here, transfers a phosphate group from ATP to an amino acid side chain of the target protein by a protein kinase. Removal of the phosphate group is catalyzed by a second enzyme, a protein phosphatase. In this example, the phosphate is added to a serine side chain; in other cases, the phosphate is instead linked to the –OH group of a threonine or a tyrosine in the protein. (B) The phosphorylation of a protein by a protein kinase can either increase or decrease the protein's activity, depending on the site of phosphorylation and the structure of the protein.

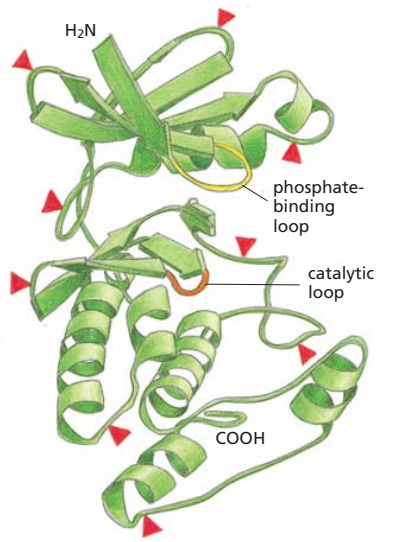



**Figure 3–65 The three-dimensional structure of a protein kinase.** Superimposed on this structure are *red arrowheads* to indicate sites where insertions of 5–100 amino acids are found in some members of the protein kinase family. These insertions are located in loops on the surface of the enzyme where other ligands interact with the protein. Thus, they distinguish different kinases and confer on them distinctive interactions with other proteins. The ATP (which donates a phosphate group) and the peptide to be phosphorylated are held in the active site, which extends between the phosphate-binding loop *(yellow)* and the catalytic loop *(orange)*. See also Figure 3–10. (Adapted from D.R. Knighton et al., *Science* 253:407–414, 1991. With permission from AAAS.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 58 of 240

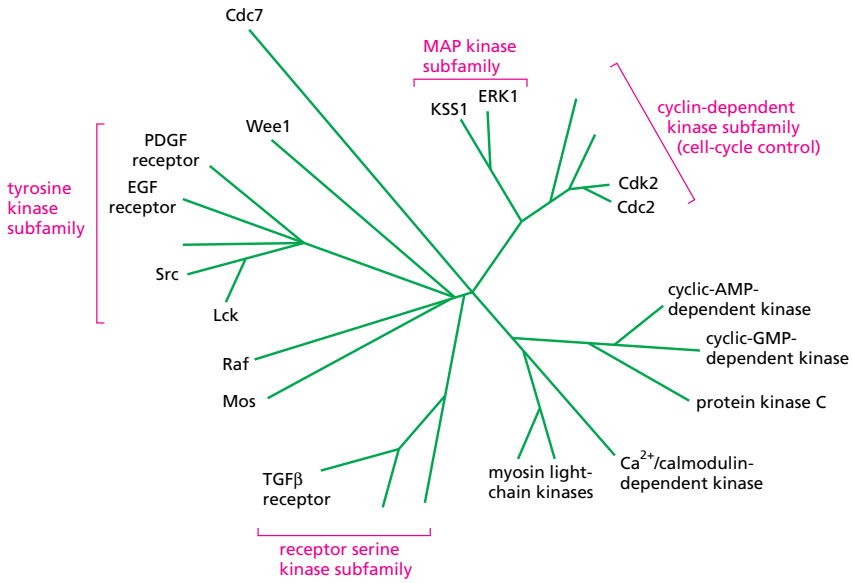



**Figure 3–66 An evolutionary tree of selected protein kinases.** Although a higher eucaryotic cell contains hundreds of such enzymes, and the human genome codes for more than 500, only some of those discussed in this book are shown.

As a result of the combined activities of protein kinases and protein phosphatases, the phosphate groups on proteins are continually turning over—being added and then rapidly removed. Such phosphorylation cycles may seem wasteful, but they are important in allowing the phosphorylated proteins to switch rapidly from one state to another: the more rapid the cycle, the faster a population of protein molecules can change its state of phosphorylation in response to a sudden change in the phosphorylation rate (see Figure 15–11). The energy required to drive this phosphorylation cycle is derived from the free energy of ATP hydrolysis, one molecule of which is consumed for each phosphorylation event.

## The Regulation of Cdk and Src Protein Kinases Shows How a Protein Can Function as a Microchip

The hundreds of different protein kinases in a eucaryotic cell are organized into complex networks of signaling pathways that help to coordinate the cell's activities, drive the cell cycle, and relay signals into the cell from the cell's environment. Many of the extracellular signals involved need to be both integrated and amplified by the cell. Individual protein kinases (and other signaling proteins) serve as input–output devices, or "microchips," in the integration process. An important part of the input to these signal processing proteins comes from the control that is exerted by phosphates added and removed from them by protein kinases and protein phosphatases, respectively.

In general, specific sets of phosphate groups serve to activate the protein, while other sets can inactivate it. A cyclin-dependent protein kinase (Cdk) provides a good example. Kinases in this class phosphorylate serines and threonines, and they are central components of the cell-cycle control system in eucaryotic cells, as discussed in detail in Chapter 17. In a vertebrate cell, individual Cdk proteins turn on and off in succession, as a cell proceeds through the different phases of its division cycle. When a particular kinase is on, it influences various aspects of cell behavior through effects on the proteins it phosphorylates.

A Cdk protein becomes active as a serine/threonine protein kinase only when it is bound to a second protein called a *cyclin*. But, as **Figure 3–67** shows, the binding of cyclin is only one of three distinct "inputs" required to activate the Cdk. In addition to cyclin binding, a phosphate must be added to a specific threonine side chain, and a phosphate elsewhere in the protein (covalently bound to a specific tyrosine side chain) must be removed. Cdk thus monitors a specific set

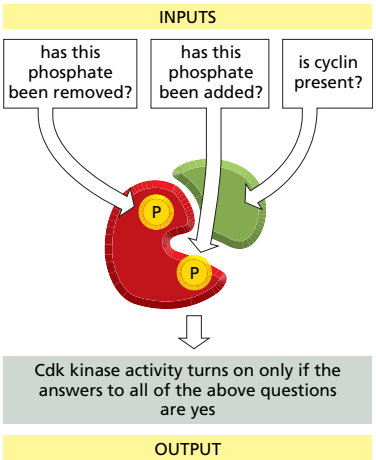



**Figure 3–67 How a Cdk protein acts as an integrating device.** <TAGA> The function of these central regulators of the cell cycle is discussed in Chapter 17.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 59 of 240




**Figure 3–68 The domain structure of the Src family of protein kinases, mapped along the amino acid sequence.** For the three-dimensional structure of Src, see Figure 3–10.

of cell components—a cyclin, a protein kinase, and a protein phosphatase—and it acts as an input–output device that turns on if, and only if, each of these components has attained its appropriate activity state. Some cyclins rise and fall in concentration in step with the cell cycle, increasing gradually in amount until they are suddenly destroyed at a particular point in the cycle. The sudden destruction of a cyclin (by targeted proteolysis) immediately shuts off its partner Cdk enzyme, and this triggers a specific step in the cell cycle.

The Src family of protein kinases (see Figure 3–10) exhibits a similar type of microchip behavior. The *Src protein* (pronounced "sarc" and named for the type of tumor, a sarcoma, that its deregulation can cause) was the first tyrosine kinase to be discovered. It is now known to be part of a subfamily of nine very similar protein kinases, which are found only in multicellular animals. As indicated by the evolutionary tree in Figure 3–66, sequence comparisons suggest that tyrosine kinases as a group were a relatively late innovation that branched off from the serine/threonine kinases, with the Src subfamily being only one subgroup of the tyrosine kinases created in this way.

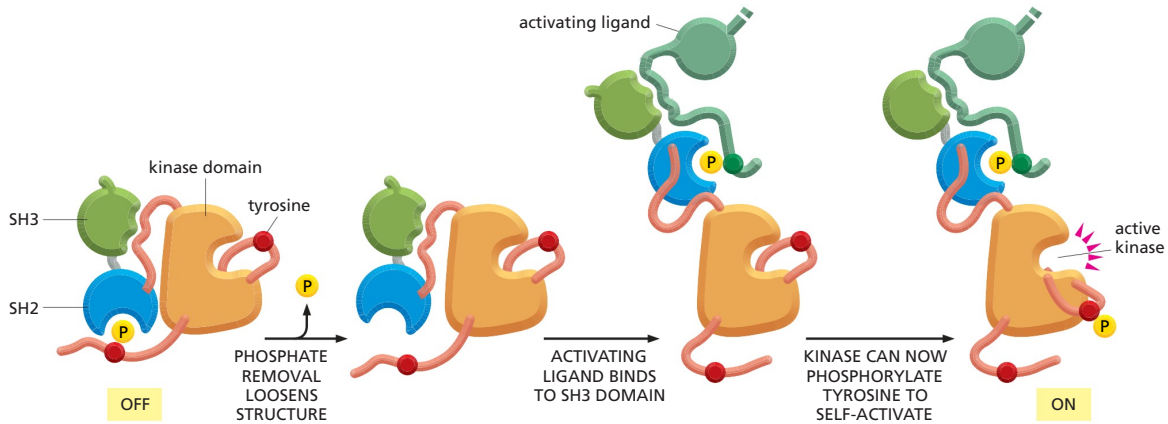
The Src protein and its relatives contain a short N-terminal region that becomes covalently linked to a strongly hydrophobic fatty acid, which holds the kinase at the cytoplasmic face of the plasma membrane. Next come two peptide-binding modules, a Src homology 3 (SH3) domain and a SH2 domain, followed by the kinase catalytic domain (**Figure 3–68**). These kinases normally exist in an inactive conformation, in which a phosphorylated tyrosine near the C-terminus is bound to the SH2 domain, and the SH3 domain is bound to an internal peptide in a way that distorts the active site of the enzyme and helps to render it inactive.

Turning the kinase on involves at least two specific inputs: removal of the C-terminal phosphate and the binding of the SH3 domain by a specific activating protein (**Figure 3–69**). Like the activation of the Cdk protein, the activation of the Src kinase signals the completion of a particular set of separate upstream events (**Figure 3–70**). Thus, both the Cdk and Src families of proteins serve as specific signal integrators, helping to generate the complex web of information-processing events that enable the cell to compute logical responses to a complex set of conditions.

## Proteins That Bind and Hydrolyze GTP Are Ubiquitous Cellular Regulators

We have described how the addition or removal of phosphate groups on a protein can be used by a cell to control the protein's activity. In the examples discussed so

**Figure 3–69 The activation of a Src-type protein kinase by two sequential events.** (Adapted from S.C. Harrison et al., *Cell* 112:737–740, 2003. With permission from Elsevier.)




LCY Biotechnology Holding, Inc.
Ex. 1008
Page 60 of 240

far, the phosphate is transferred from an ATP molecule to an amino acid side chain of the protein in a reaction catalyzed by a specific protein kinase. Eucaryotic cells also have another way to control protein activity by phosphate addition and removal. In this case, the phosphate is not attached directly to the protein; instead, it is a part of the guanine nucleotide GTP, which binds very tightly to the protein. In general, proteins regulated in this way are in their active conformations with GTP bound. The loss of a phosphate group occurs when the bound GTP is hydrolyzed to GDP in a reaction catalyzed by the protein itself, and in its GDP-bound state the protein is inactive. In this way, GTP-binding proteins act as on–off switches whose activity is determined by the presence or absence of an additional phosphate on a bound GDP molecule (**Figure 3–71**).

**GTP-binding proteins** (also called **GTPases** because of the GTP hydrolysis they catalyze) comprise a large family of proteins that all contain variations on the same GTP-binding globular domain. When the tightly bound GTP is hydrolyzed to GDP, this domain undergoes a conformational change that inactivates it. The three-dimensional structure of a prototypical member of this family, the monomeric GTPase called Ras, is shown in **Figure 3–72**.

The *Ras protein* has an important role in cell signaling (discussed in Chapter 15). In its GTP-bound form, it is active and stimulates a cascade of protein phosphorylations in the cell. Most of the time, however, the protein is in its inactive, GDP-bound form. It becomes active when it exchanges its GDP for a GTP molecule in response to extracellular signals, such as growth factors, that bind to receptors in the plasma membrane (see Figure 15–58).

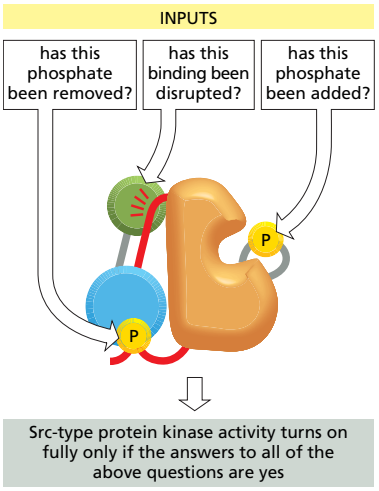



**Figure 3–70 How a Src-type protein kinase acts as an integrating device.** The disruption of the SH3 domain interaction *(green)* involves replacing its binding to the indicated *red* linker region with a tighter interaction with an activating ligand, as illustrated in Figure 3–69.

## Regulatory Proteins Control the Activity of GTP-Binding Proteins by Determining Whether GTP or GDP Is Bound

GTP-binding proteins are controlled by regulatory proteins that determine whether GTP or GDP is bound, just as phosphorylated proteins are turned on and off by protein kinases and protein phosphatases. Thus, Ras is inactivated by a *GTPase-activating protein (GAP)*, which binds to the Ras protein and induces it to hydrolyze its bound GTP molecule to GDP—which remains tightly bound—and inorganic phosphate ($P_i$), which is rapidly released. The Ras protein stays in its inactive, GDP-bound conformation until it encounters a *guanine nucleotide exchange factor (GEF)*, which binds to GDP-Ras and causes it to release its GDP. Because the empty nucleotide-binding site is immediately filled by a GTP molecule (GTP is present in large excess over GDP in cells), the GEF activates Ras by *indirectly* adding back the phosphate removed by GTP hydrolysis. Thus, in a sense, the roles of GAP and GEF are analogous to those of a protein phosphatase and a protein kinase, respectively (**Figure 3–73**).

## Large Protein Movements Can Be Generated From Small Ones

The Ras protein belongs to a large superfamily of *monomeric GTPases*, each of which consists of a single GTP-binding domain of about 200 amino acids. Over the course of evolution, this domain has also become joined to larger proteins with additional domains, creating a large family of GTP-binding proteins. Family members include the receptor-associated trimeric G proteins involved in cell signaling (discussed in Chapter 15), proteins regulating the traffic of vesicles between intracellular compartments (discussed in Chapter 13), and proteins that bind to transfer RNA and are required as assembly factors for protein

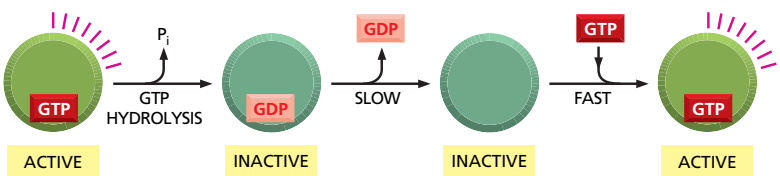



**Figure 3–71 GTP-binding proteins as molecular switches.** The activity of a GTP-binding protein (also called a GTPase) generally requires the presence of a tightly bound GTP molecule (switch "on"). Hydrolysis of this GTP molecule produces GDP and inorganic phosphate ($P_i$), and it causes the protein to convert to a different, usually inactive, conformation (switch "off"). As shown here, resetting the switch requires the tightly bound GDP to dissociate, a slow step that is greatly accelerated by specific signals; once the GDP has dissociated, a molecule of GTP is quickly rebound.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 61 of 240

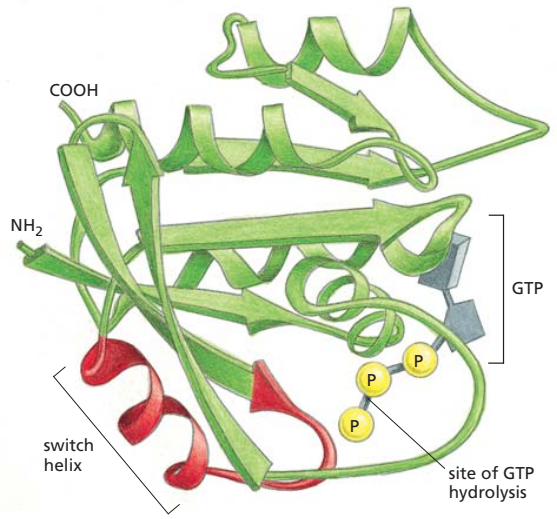



**Figure 3–72 The structure of the Ras protein in its GTP-bound form.** <GAAC> This monomeric GTPase illustrates the structure of a GTP-binding domain, which is present in a large family of GTP-binding proteins. The *red* regions change their conformation when the GTP molecule is hydrolyzed to GDP and inorganic phosphate by the protein; the GDP remains bound to the protein, while the inorganic phosphate is released. The special role of the "switch helix" in proteins related to Ras is explained next (see Figure 3–75).

synthesis on the ribosome (discussed in Chapter 6). In each case, an important biological activity is controlled by a change in the protein's conformation that is caused by GTP hydrolysis in a Ras-like domain.

The *EF-Tu protein* provides a good example of how this family of proteins works. EF-Tu is an abundant molecule that serves as an elongation factor (hence the EF) in protein synthesis, loading each aminoacyl tRNA molecule onto the ribosome. The tRNA molecule forms a tight complex with the GTP-bound form of EF-Tu (**Figure 3–74**). In this complex, the amino acid attached to the tRNA is improperly positioned for protein synthesis. The tRNA can transfer its amino acid only after the GTP bound to EF-Tu is hydrolyzed on the ribosome, allowing the EF-Tu to dissociate. Since the GTP hydrolysis is triggered by a proper fit of the tRNA to the mRNA molecule on the ribosome, the EF-Tu serves as a factor that discriminates between correct and incorrect mRNA–tRNA pairings (see Figure 6–67 for a further discussion of this function of EF-Tu).

By comparing the three-dimensional structure of EF-Tu in its GTP-bound and GDP-bound forms, we can see how the repositioning of the tRNA occurs. The dissociation of the inorganic phosphate group ($P_i$), which follows the reaction GTP → GDP + $P_i$, causes a shift of a few tenths of a nanometer at the GTP-binding site, just as it does in the Ras protein. This tiny movement, equivalent to

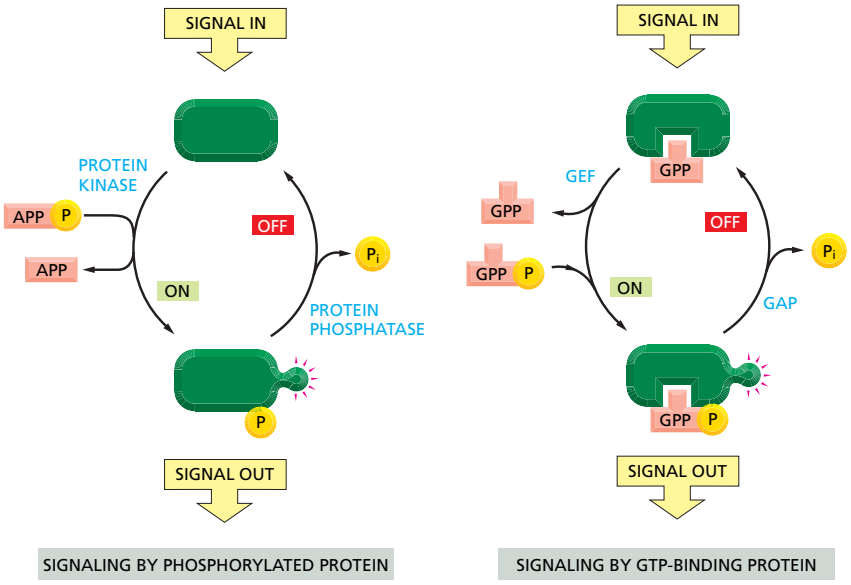



**Figure 3–73 A comparison of the two major intracellular signaling mechanisms in eucaryotic cells.** In both cases, a signaling protein is activated by the addition of a phosphate group and inactivated by the removal of this phosphate. To emphasize the similarities in the two pathways, ATP and GTP are drawn as APPP and GPPP, and ADP and GDP as APP and GPP, respectively. As shown in Figure 3–64, the addition of a phosphate to a protein can also be inhibitory.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 62 of 240

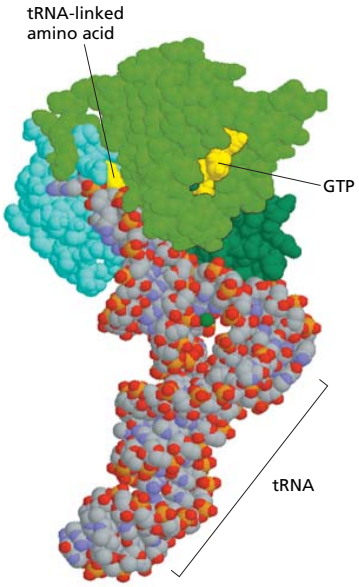



**Figure 3–74 An aminoacyl tRNA molecule bound to EF-Tu.** The three domains of the EF-Tu protein are colored differently, to match Figure 3–75. This is a bacterial protein; however, a very similar protein exists in eucaryotes, where it is called EF-1. (Coordinates determined by P. Nissen et al., *Science* 270:1464–1472, 1995. With permission from AAAS.)

a few times the diameter of a hydrogen atom, causes a conformational change to propagate along a crucial piece of α helix, called the *switch helix*, in the Ras-like domain of the protein. The switch helix seems to serve as a latch that adheres to a specific site in another domain of the molecule, holding the protein in a "shut" conformation. The conformational change triggered by GTP hydrolysis causes the switch helix to detach, allowing separate domains of the protein to swing apart, through a distance of about 4 nm. This releases the bound tRNA molecule, allowing its attached amino acid to be used (**Figure 3–75**).

Notice in this example how cells have exploited a simple chemical change that occurs on the surface of a small protein domain to create a movement 50 times larger. Dramatic shape changes of this type also cause the very large movements that occur in motor proteins, as we discuss next.

## Motor Proteins Produce Large Movements in Cells

We have seen that conformational changes in proteins have a central role in enzyme regulation and cell signaling. We now discuss proteins whose major function is to move other molecules. These **motor proteins** generate the forces responsible for muscle contraction and the crawling and swimming of cells. Motor proteins also power smaller-scale intracellular movements: they help to move chromosomes to opposite ends of the cell during mitosis (discussed in Chapter 17), to move organelles along molecular tracks within the cell (discussed

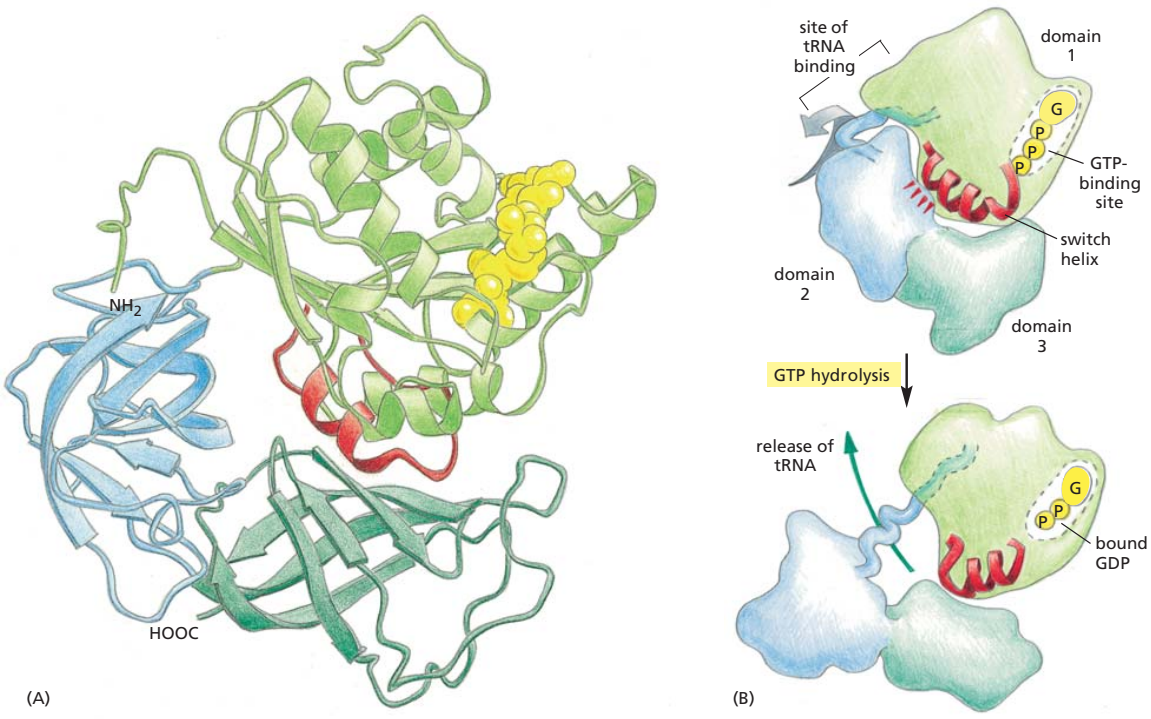



**Figure 3–75 The large conformational change in EF-Tu caused by GTP hydrolysis.** <GTAA> (A) The three-dimensional structure of EF-Tu with GTP bound. The domain at the top has a structure similar to the Ras protein, and its *red* α helix is the switch helix, which moves after GTP hydrolysis. (B) The change in the conformation of the switch helix in domain 1 causes domains 2 and 3 to rotate as a single unit by about 90° toward the viewer, which releases the tRNA that was shown bound to this structure in Figure 3–74. (A, adapted from H. Berchtold et al., *Nature* 365:126–132, 1993. With permission from Macmillan Publishers Ltd. B, courtesy of Mathias Sprinzl and Rolf Hilgenfeld.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 63 of 240

in Chapter 16), and to move enzymes along a DNA strand during the synthesis of a new DNA molecule (discussed in Chapter 5). All these fundamental processes depend on proteins with moving parts that operate as force-generating machines.

How do these machines work? In other words, how do cells use shape changes in proteins to generate directed movements? If, for example, a protein is required to walk along a narrow thread such as a DNA molecule, it can do this by undergoing a series of conformational changes, such as those shown in **Figure 3–76**. But with nothing to drive these changes in an orderly sequence, they are perfectly reversible, and the protein can only wander randomly back and forth along the thread. We can look at this situation in another way. Since the directional movement of a protein does work, the laws of thermodynamics (discussed in Chapter 2) demand that such movement use free energy from some other source (otherwise the protein could be used to make a perpetual motion machine). Therefore, without an input of energy, the protein molecule can only wander aimlessly.

How can the cell make such a series of conformational changes unidirectional? To force the entire cycle to proceed in one direction, it is enough to make any one of the changes in shape irreversible. Most proteins that are able to walk in one direction for long distances achieve this motion by coupling one of the conformational changes to the hydrolysis of an ATP molecule bound to the protein. The mechanism is similar to the one just discussed that drives allosteric protein shape changes by GTP hydrolysis. Because ATP (or GTP) hydrolysis releases a great deal of free energy, it is very unlikely that the nucleotide-binding protein will undergo the reverse shape change needed for moving backward—since this would require that it also reverse the ATP hydrolysis by adding a phosphate molecule to ADP to form ATP.

In the model shown in **Figure 3–77**, ATP binding shifts a motor protein from conformation 1 to conformation 2. The bound ATP is then hydrolyzed to produce ADP and inorganic phosphate ($P_i$), causing a change from conformation 2 to conformation 3. Finally, the release of the bound ADP and $P_i$ drives the protein back to conformation 1. Because the energy provided by ATP hydrolysis drives the transition 2 → 3, this series of conformational changes is effectively irreversible. Thus, the entire cycle goes in only one direction, causing the protein molecule to walk continuously to the right in this example.

Many motor proteins generate directional movement in this general way, including the muscle motor protein *myosin*, which walks along actin filaments to generate muscle contraction, and the *kinesin* proteins that walk along microtubules (both discussed in Chapter 16). These movements can be rapid: some of the motor proteins involved in DNA replication (the DNA helicases) propel themselves along a DNA strand at rates as high as 1000 nucleotides per second.

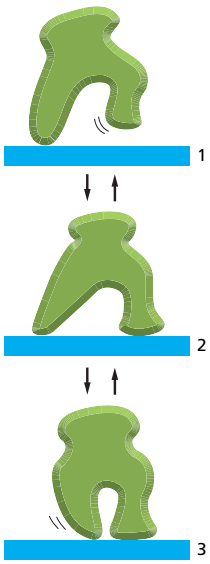



**Figure 3–76 An allosteric "walking" protein.** Although its three different conformations allow it to wander randomly back and forth while bound to a thread or a filament, the protein cannot move uniformly in a single direction.

## Membrane-Bound Transporters Harness Energy to Pump Molecules Through Membranes

We have thus far seen how allosteric proteins can act as microchips (Cdk and Src kinases), as assembly factors (EF-Tu), and as generators of mechanical force and motion (motor proteins). Allosteric proteins can also harness energy derived from ATP hydrolysis, ion gradients, or electron transport processes to pump specific ions or small molecules across a membrane. We consider one example here; others will be discussed in Chapter 11.

The ABC transporters constitute an important class of membrane-bound pump proteins. In humans at least 48 different genes encode them. These transporters mostly function to export hydrophobic molecules from the cytoplasm,

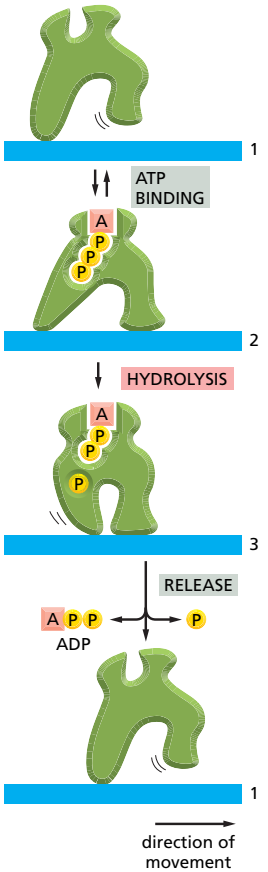



**Figure 3–77 An allosteric motor protein.** The transition between three different conformations includes a step driven by the hydrolysis of a bound ATP molecule, and this makes the entire cycle essentially irreversible. By repeated cycles, the protein therefore moves continuously to the right along the thread.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 64 of 240

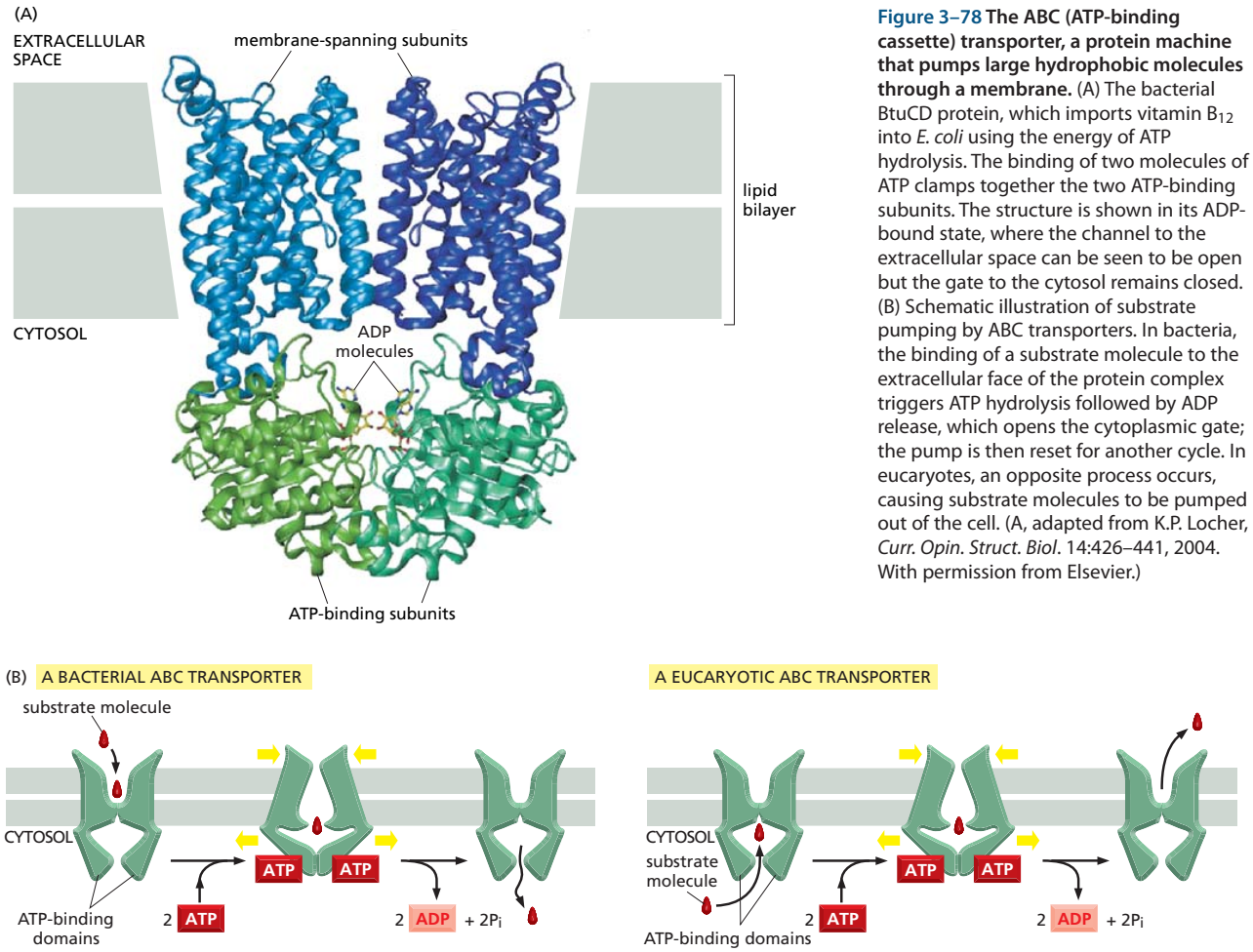



**Figure 3–78 The ABC (ATP-binding cassette) transporter, a protein machine that pumps large hydrophobic molecules through a membrane.** (A) The bacterial BtuCD protein, which imports vitamin $B_{12}$ into *E. coli* using the energy of ATP hydrolysis. The binding of two molecules of ATP clamps together the two ATP-binding subunits. The structure is shown in its ADP-bound state, where the channel to the extracellular space can be seen to be open but the gate to the cytosol remains closed. (B) Schematic illustration of substrate pumping by ABC transporters. In bacteria, the binding of a substrate molecule to the extracellular face of the protein complex triggers ATP hydrolysis followed by ADP release, which opens the cytoplasmic gate; the pump is then reset for another cycle. In eucaryotes, an opposite process occurs, causing substrate molecules to be pumped out of the cell. (A, adapted from K.P. Locher, *Curr. Opin. Struct. Biol.* 14:426–441, 2004. With permission from Elsevier.)

serving to remove toxic molecules at the mucosal surface of the intestinal tract, for example, or at the blood–brain barrier. The study of ABC transporters is of intense interest in clinical medicine, because the overproduction of proteins in this class contributes to the resistance of tumor cells to chemotherapeutic drugs. And in bacteria, the same type of proteins primarily function to import essential nutrients into the cell.

The ABC transporter is a tetramer, with a pair of membrane-spanning subunits linked to a pair of ATP binding subunits located just below the plasma membrane (**Figure 3–78**A). As in other examples we have discussed, the hydrolysis of the bound ATP molecules drives conformational changes in the protein, transmitting forces that cause the membrane-spanning subunits to move their bound molecules across the lipid bilayer (Figure 3–78B).

Humans have invented many different types of mechanical pumps, and it should not be surprising that cells also contain membrane-bound pumps that function in other ways. Among the most notable are the rotary pumps that couple the hydrolysis of ATP to the transport of $H^+$ ions (protons). These pumps resemble miniature turbines, and they are used to acidify the interior of lysosomes and other eucaryotic organelles. Like other ion pumps that create ion gradients, they can function in reverse to catalyze the reaction ADP + $P_i \rightarrow$ ATP, if the gradient across their membrane of the ion that they transport is steep enough.

One such pump, the ATP synthase, harnesses a gradient of proton concentration produced by electron transport processes to produce most of the ATP used in the living world. This ubiquitous pump has a central role in energy conversion, and we shall discuss its three-dimensional structure and mechanism in Chapter 14.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 65 of 240

## Proteins Often Form Large Complexes That Function as Protein Machines <ACTT> <ATCG>

Large proteins formed from many domains are able to perform more elaborate functions than small, single-domain proteins. But large protein assemblies formed from many protein molecules perform the most impressive tasks. Now that it is possible to reconstruct most biological processes in cell-free systems in the laboratory, it is clear that each of the central processes in a cell—such as DNA replication, protein synthesis, vesicle budding, or transmembrane signaling—is catalyzed by a highly coordinated, linked set of 10 or more proteins. In most such *protein machines*, an energetically favorable reaction such as the hydrolysis of bound nucleoside triphosphates (ATP or GTP) drives an ordered series of conformational changes in one or more of the individual protein subunits, enabling the ensemble of proteins to move coordinately. In this way, each enzyme can be moved directly into position, as the machine catalyzes successive reactions in a series. This is what occurs, for example, in protein synthesis on a ribosome (discussed in Chapter 6)—or in DNA replication, where a large multiprotein complex moves rapidly along the DNA (discussed in Chapter 5).

Cells have evolved protein machines for the same reason that humans have invented mechanical and electronic machines. For accomplishing almost any task, manipulations that are spatially and temporally coordinated through linked processes are much more efficient than the use of individual tools.

## Protein Machines with Interchangeable Parts Make Efficient Use of Genetic Information

To probe more deeply into the nature of protein machines, we shall consider a relatively simple one: the **SCF ubiquitin ligase**. This protein complex binds different "target proteins" at different times in the cell cycle, and it covalently adds multiubiquitin polypeptide chains to these proteins. Its C-shaped structure is formed from five protein subunits, the largest of which is a molecule that serves as a *scaffold protein* on which the rest of the structure is built. The structure underlies a remarkable mechanism (**Figure 3–79**). At one end of the C is an E2 ubiquitin-conjugating enzyme. At the other end is a substrate-binding arm, a subunit known as an *F-box protein*. These two subunits are separated by a gap of about 5 nm. When this protein complex is activated, the F-box protein binds to a specific site on a target protein, positioning the protein in the gap so that some of its lysine side chains contact the ubiquitin-conjugating enzyme. This enzyme can then catalyze the repeated addition of a ubiquitin polypeptide to these lysines (see Figure 3–79C), producing a polyubiquitin chain that marks the target protein for rapid destruction in a proteasome (see p. 393).

In this manner, specific proteins are targeted for rapid destruction in response to specific signals, thereby helping to drive the cell cycle (discussed in Chapter 17). The timing of the destruction often involves creating a specific pattern of phosphorylation on the target protein that is required for its recognition by the F-box subunit. It also requires the activation of an SCF ubiquitin ligase that carries the appropriate substrate-binding arm. Many of these arms (the F-box subunits) are interchangeable in the protein complex (see Figure 3–79B), and there are more than 70 human genes that encode them.

As emphasized previously, once a successful protein has evolved, its genetic information tends to be duplicated to produce a family of related proteins. Thus, for example, not only are there many F-box proteins—making possible the recognition of different sets of target proteins—but there is also a family of scaffolds (known as cullins) that give rise to a family of SCF-like ubiquitin ligases.

The pressure on organisms to minimize the number of genes (see p. 265) presumably helps to explain why RNA splicing is so prevalent in higher eucaryotes, allowing multiple related proteins to be synthesized from a single gene (discussed in Chapter 6). A protein machine like the SCF ubiquitin ligase, with its interchangeable parts, likewise makes economical use of the genetic information

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 66 of 240

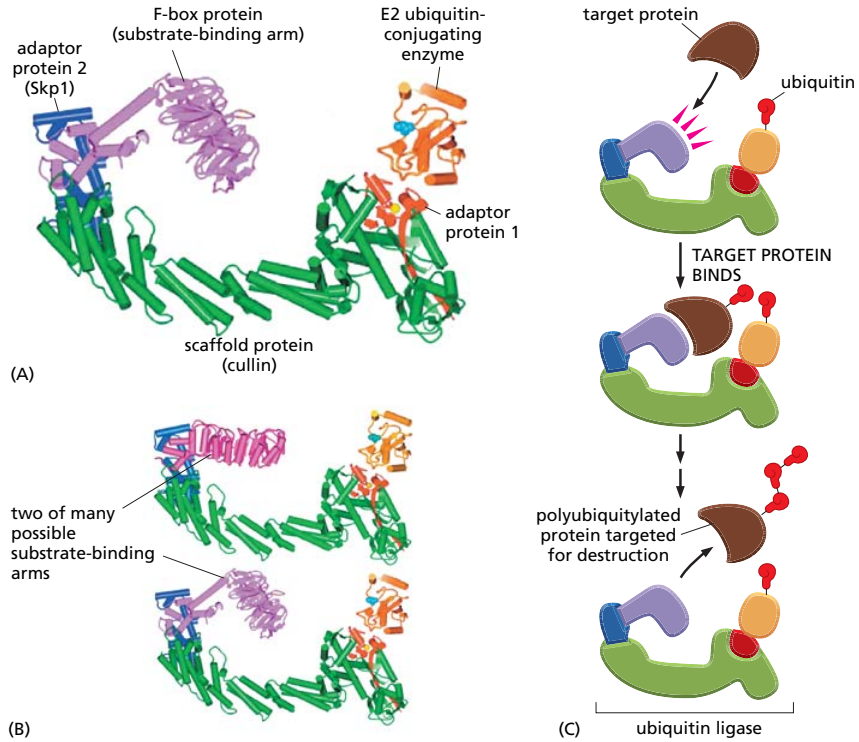



**Figure 3–79 The structure and mode of action of a SCF ubiquitin ligase.** (A) The structure of the five-protein complex that includes an E2 ubiquitin ligase. The protein denoted here as adapter protein 1 is the Rbx1/Hrt1 protein, adaptor protein 2 is the Skp1 protein, and the cullin is the Cul1 protein. (B) Comparison of the same complex with two different substrate-binding arms, the F-box proteins Skp2 *(top)* and β-trCP1 *(bottom),* respectively. (C) The binding and ubiquitylation of a target protein by the SCF ubiquitin ligase. If, as indicated, a chain of ubiquitin molecules is added to the same lysine of the target protein, that protein is marked for rapid destruction by the proteasome. (A and B, adapted from G. Wu et al., *Mol. Cell* 11:1445–1456, 2003. With permission from Elsevier.)

in cells, inasmuch as new functions can evolve for the entire complex simply by producing an alternative version of one of its subunits.

## The Activation of Protein Machines Often Involves Positioning Them at Specific Sites

As scientists have learned more of the details of cell biology, they have recognized increasing degrees of sophistication in cell chemistry. Thus, not only do we now know that protein machines play a predominant role, but it has recently become clear that most of these machines form at specific sites in the cell, being activated only where and when they are needed. Using fluorescent, GFP-tagged fusion proteins in living cells (see p. 593), cell biologists are able to follow the repositioning of individual proteins that occurs in response to specific signals. Thus, when certain extracellular signaling molecules bind to receptor proteins in the plasma membrane, they often recruit a set of other proteins to the inside surface of the plasma membrane to form protein machines that pass the signal on. As an example, **Figure 3–80**A illustrates the rapid movement of a protein kinase C (PKC) enzyme to a complex in the plasma membrane, where it associates with specific substrate proteins that it phosphorylates.

There are more than 10 distinct PKC enzymes in human cells, which differ both in their mode of regulation and in their functions. When activated, these enzymes move from the cytoplasm to different intracellular locations, forming specific complexes with other proteins that allow them to phosphorylate different protein substrates (Figure 3–80B). The SCF ubiquitin ligases can also move to specific sites of function at appropriate times. As will be explained when we discuss cell signaling in Chapter 15, the mechanisms frequently involve protein phosphorylation, as well as **scaffold proteins** that link together a set of activating, inhibiting, adaptor, and substrate proteins at a specific location in a cell.

This general phenomenon is known as *induced proximity*, and it explains the otherwise puzzling observation that slightly different forms of enzymes with the same catalytic site will often have very different biological functions. Cells change the locations of their proteins by covalently modifying them in a variety of different ways, as part of a "regulatory code" to be described next.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 67 of 240

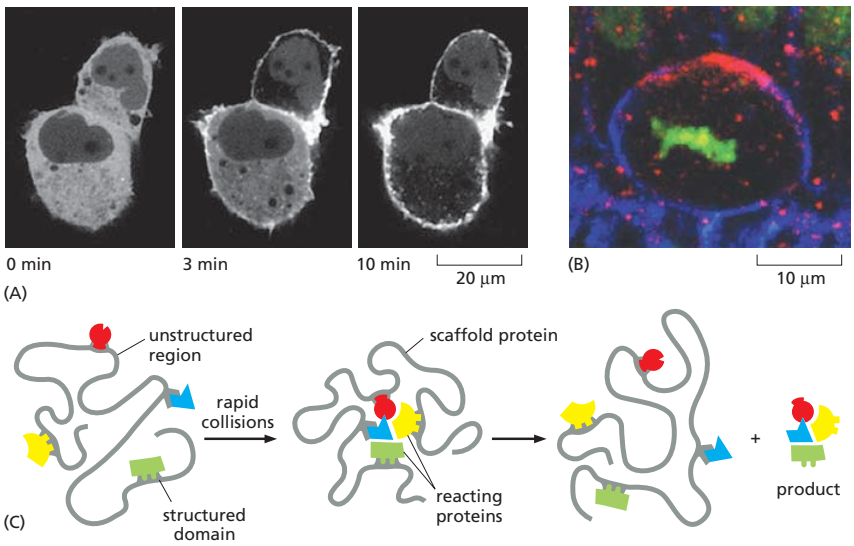



**Figure 3–80 The assembly of protein machines at specific sites in a cell.** (A) In response to a signal (here a phorbol ester), the gamma subspecies of protein kinase C moves rapidly from the cytosol to the plasma membrane. The protein kinase is fluorescent in these living cells because an engineered gene inside the cell encodes a fusion protein that links the kinase to green fluorescent protein (GFP). (B) The specific association of a different subspecies of protein kinase C (aPKC) with the apical tip of a differentiating neuroblast in an early *Drosophila* embryo. The kinase is stained *red*, and the cell nucleus *green*. (C) Diagram illustrating how a simple proximity created by scaffold proteins can greatly speed reactions in a cell. In this example, long unstructured regions of polypeptide chain in a large scaffold protein connect a series of structured domains that bind a set of reacting proteins. The unstructured regions serve as flexible "tethers" that greatly speed reaction rates by causing a rapid, random collision of all of the proteins that are bound to the scaffold. (For a simple example of tethering, see Figure 16–38.) (A, from N. Sakai et al, *J. Cell Biol.* 139:1465–1476, 1997. With permission from The Rockefeller University Press. B, courtesy of Andreas Wodarz, Institute of Genetics, University of Düsseldorf, Germany.)

These modifications create sites on proteins that bind them to particular scaffold proteins, thereby clustering the proteins required for particular reactions in specific regions of the cell. Most biological reactions are catalyzed by sets of 5 or more proteins, and such a clustering of proteins is often required for the reaction to occur. Scaffolds thereby allow cells to compartmentalize reactions even in the absence of membranes. Although only recently recognized as a widespread phenomenon, this type of clustering is particularly obvious in the cell nucleus (see Figure 4–69).

Many scaffolds appear to be quite different from the cullin illustrated previously in Figure 3–79: rather than holding their bound proteins in precise positions relative to each other, the interacting proteins are linked by unstructured regions of polypeptide chain. This tethers the proteins together, causing them to collide frequently with each other in random orientations—some of which will lead to a productive reaction (Figure 3–80C). In essence, this mechanism greatly speeds reactions by creating a very high local concentration of the reacting species. For this reason, the use of scaffold proteins represents an especially versatile way of controlling cell chemistry (see also Figure 15–61).

## Many Proteins Are Controlled by Multisite Covalent Modification

We have thus far described only one type of posttranslational modification of proteins—that in which a phosphate is covalently attached to an amino acid side chain (see Figure 3–64). But a large number of other such modifications also occur, more than 200 distinct types being known. To give a sense of the variety, **Table 3–3** presents a subset of modifying groups with known regulatory roles. As

**Table 3–3 Some Molecules Covalently Attached to Proteins Regulate Protein Function**

| MODIFYING GROUP | SOME PROMINENT FUNCTIONS |
|---|---|
| **Phosphate** on Ser, Thr, or Tyr | Drives the assembly of a protein into larger complexes (see Figure 15–19). |
| **Methyl** on Lys | Helps to creates histone code in chromatin through forming either mono-, di-, or tri-methyl lysine (see Figure 4–38). |
| **Acetyl** on Lys | Helps to creates histone code in chromatin (see Figure 4–38). |
| **Palmityl** group on Cys | This fatty acid addition drives protein association with membranes (see Figure 10–20). |
| ***N*-acetylglucosamine** on Ser or Thr | Controls enzyme activity and gene expression in glucose homeostasis. |
| **Ubiquitin** on Lys | Monoubiquitin addition regulates the transport of membrane proteins in vesicles (see Figure 13–58). |
| | A polyubiquitin chain targets a protein for degradation (see Figure 3–79). |

Ubiquitin is a 76 amino acid polypeptide; there are at least 10 other ubiquitin-related proteins, such as SUMO, that modify proteins in similar ways.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 68 of 240

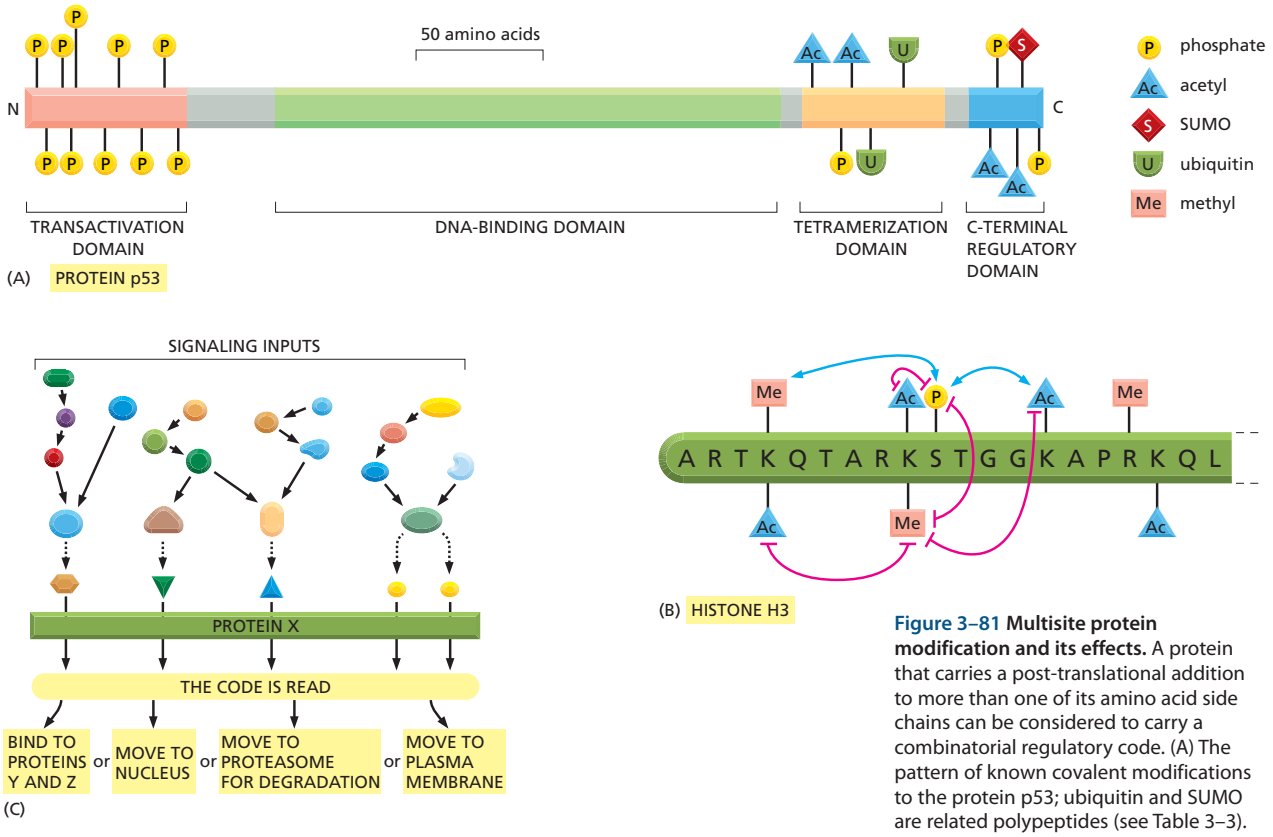



**Figure 3–81 Multisite protein modification and its effects.** A protein that carries a post-translational addition to more than one of its amino acid side chains can be considered to carry a combinatorial regulatory code. (A) The pattern of known covalent modifications to the protein p53; ubiquitin and SUMO are related polypeptides (see Table 3–3). (B) The possible modifications on the first 20 amino acids at the N-terminus of histone H3, showing not only their locations but also their activating *(blue)* and inhibiting *(red)* effects on the addition of neighboring covalent modifications. In addition to the effects shown, the acetylation and methylation of a lysine are mutually exclusive reactions (see Figure 4–38). (C) Diagram showing the general manner in which multisite modifications are added to (and removed from) a protein through signaling networks, and how the resulting combinatorial regulatory code on the protein is read to alter its behavior in the cell.

in phosphate addition, these groups are added and then removed from proteins according to the needs of the cell.

A large number of proteins are now known to be modified on more than one amino acid side chain, with different regulatory events producing a different pattern of such modifications. A striking example is the protein p53, which plays a central part in controlling a cell's response to adverse circumstances (see p. 1105). Through one of four different types of molecular additions, this protein can be modified at 20 different sites (**Figure 3–81**A). Because an enormous number of different combinations of these 20 modifications are possible, the protein's behavior can in principle be altered in a huge number of ways. Moreover, the pattern of modifications on a protein can determine its susceptibility to further modification, as illustrated by histone H3 in Figure 3–81B.

Cell biologists have only recently come to recognize that each protein's set of covalent modifications constitutes an important *combinatorial regulatory code.* As specific modifying groups are added to or removed from a protein, this code causes a different set of protein behaviors—changing the activity or stability of the protein, its binding partners, and its specific location within the cell (Figure 3–81C). This helps the cell respond rapidly and with great versatility to changes in its condition or environment.

## A Complex Network of Protein Interactions Underlies Cell Function

There are many challenges facing cell biologists in this "post-genome" era when complete genome sequences are known. One is the need to dissect and reconstruct each one of the thousands of protein machines that exist in an organism such as ourselves. To understand these remarkable protein complexes, each must be reconstituted from its purified protein parts, so that we can study its detailed mode of operation under controlled conditions in a test tube, free from

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 69 of 240

all other cell components. This alone is a massive task. But we now know that each of these subcomponents of a cell also interacts with other sets of macromolecules, creating a large network of protein–protein and protein–nucleic acid interactions throughout the cell. To understand the cell, therefore, we need to analyze most of these other interactions as well.

We can gain some idea of the complexity of intracellular protein networks from a particularly well-studied example described in Chapter 16: the many dozens of proteins that interact with the actin cytoskeleton in the yeast *Saccharomyces cerevisiae* (see Figure 16–18). The extent of such protein–protein interactions can also be estimated more generally. An enormous amount of valuable information is now freely available in protein databases on the Internet: tens of thousands of three-dimensional protein structures plus tens of millions of protein sequences derived from the nucleotide sequences of genes. Scientists have been developing new methods for mining this great resource to increase our understanding of cells. In particular, computer-based bioinformatics tools are being combined with robotics and microarray technologies (see p. 574) to allow thousands of proteins to be investigated in a single set of experiments. **Proteomics** is a term that is often used to describe such research focused on the large-scale analysis of proteins, analogous to the term *genomics* describing the large-scale analysis of DNA sequences and genes.

Biologists use two different large-scale methods to map the direct binding interactions between the many different proteins in a cell. The initial method of choice was based on genetics: through an ingenious technique known as the yeast two-hybrid screen (see Figure 8–24), tens of thousands of interactions between thousands of proteins have been mapped in yeast, a nematode, and the fruit fly *Drosophila*. More recently, a biochemical method based on affinity tagging and mass spectroscopy has gained favor (discussed in Chapter 8), because it appears to produce fewer spurious results. The results of these and other analyses that predict protein binding interactions have been tabulated and organized in Internet databases. This allows a cell biologist studying a small set of proteins to readily discover which other proteins in the same cell are thought to bind to, and thus interact with, that set of proteins. When displayed graphically as a *protein interaction map*, each protein is represented by a box or dot in a two-dimensional network, with a straight line connecting those proteins that have been found to bind to each other.

When hundreds or thousands of proteins are displayed on the same map, the network diagram becomes bewilderingly complicated, serving to illustrate how much more we have to learn before we can claim to really understand the cell. Much more useful are small subsections of these maps, centered on a few proteins of interest. Thus, **Figure 3–82** shows a network of protein–protein interactions for the five proteins that form the SCF ubiquitin ligase in a yeast cell (see Figure 3–79). Four of the subunits of this ligase are located at the bottom right of Figure 3–82. The remaining subunit, the F-box protein that serves as its substrate-binding arm, appears as a set of 15 different gene products that bind to adaptor protein 2 (the Skp1 protein). Along the top and left of the figure are sets of additional protein interactions marked with *yellow* and *green* shading: as indicated, these protein sets function at the origin of DNA replication, in cell cycle regulation, in methionine synthesis, in the kinetochore, and in vacuolar $H^+$-ATPase assembly. We shall use this figure to explain how such protein interaction maps are used, and what they do and do not mean.

1. Protein interaction maps are useful for identifying the likely function of previously uncharacterized proteins. Examples are the products of the genes that have thus far only been inferred to exist from the yeast genome sequence, which are the six proteins in the figure that lack a simple three-letter abbreviation (*white letters* beginning with Y). One, the product of so-called *open reading frame* YDR196C, is located in the origin of replication group, and it is therefore likely to have a role in starting new replication forks. The remaining five in this diagram are F-box proteins that bind to Skp1; these are therefore likely to function as part of the ubiquitin ligase, serving as substrate-binding arms that recognize different target proteins.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 70 of 240

However, as we discuss next, neither assignment can be considered certain without additional data.

2. Protein interaction networks need to be interpreted with caution because, as a result of evolution making efficient use of each organism's genetic information, the same protein can be used as part of two different protein complexes that have different types of functions. Thus, although protein A binds to protein B and protein B binds to protein C, proteins A and C need not function in the same process. For example, we know from detailed biochemical studies that the functions of Skp1 in the kinetochore and in vacuolar $H^+$-ATPase assembly *(yellow shading)* are separate from its function in the SCF ubiquitin ligase. In fact, only the remaining three functions of Skp1 illustrated in the diagram—methionine synthesis, cell cycle regulation, and origin of replication *(green shading)*—involve ubiquitylation.
3. In cross-species comparisons, those proteins displaying similar patterns of interactions in the two protein interaction maps are likely to have the same function in the cell. Thus, as scientists generate more and more highly detailed maps for multiple organisms, the results will become increasingly useful for inferring protein function. These map comparisons are a particularly powerful tool for deciphering the functions of human proteins. There is a vast amount of direct information about protein function that can be obtained from genetic engineering, mutational, and

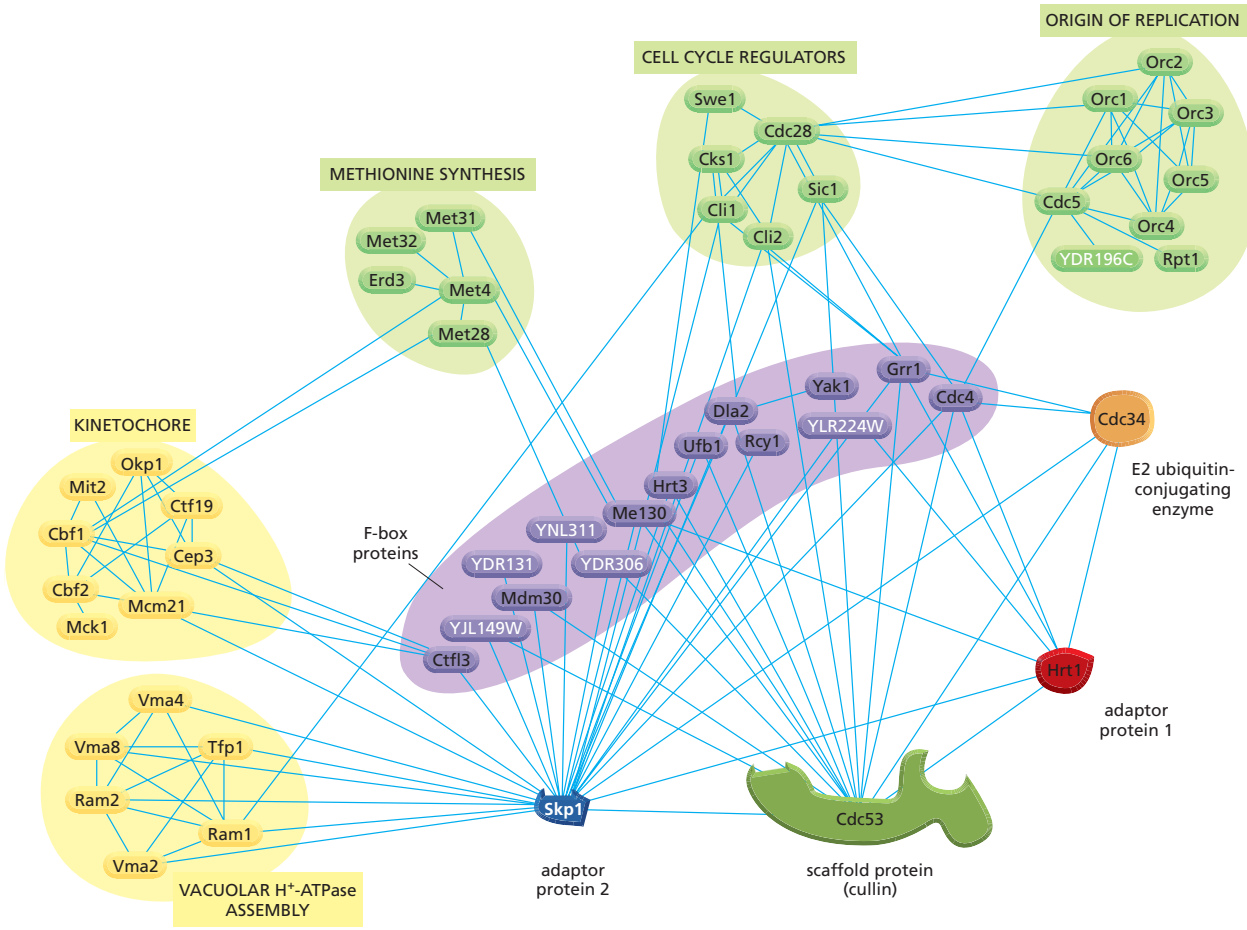



**Figure 3–82 A map of some protein– protein interactions of the SCF ubiquitin ligase and other proteins in the yeast *S. cerevisiae*.** The symbols and/or colors used for the 5 proteins of the ligase are those in Figure 3–79. Note that 15 different F-box proteins are shown (*purple*); those with *white* lettering (beginning with Y) are only known from the genome sequence as open reading frames. For additional details, see text. (Courtesy of Peter Bowers and David Eisenberg, UCLA-DOE Institute for Genomics and Proteomics, UCLA.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 71 of 240

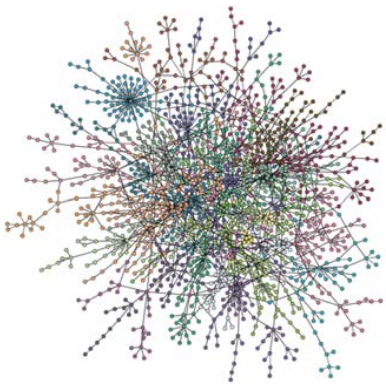


**Figure 3–83 A network of protein-binding interactions in a yeast cell.** Each line connecting a pair of dots (proteins) indicates a protein–protein interaction. (From A. Guimerá and M. Sales–Pardo, *Mol. Syst. Biol.* 2:42, 2006. With permission from Macmillan Publishers Ltd.)

genetic analyses in model organisms—such as yeast, worms, and flies—that is not available in humans.

The available data suggest that a typical protein in a human cell may interact with between 5 and 15 different partners. Often, each of the different domains in a multidomain protein binds to a different set of partners; in fact, we can speculate that the unusually extensive multidomain structures observed for human proteins may have evolved to facilitate these interactions. Given the enormous complexity of the interacting networks of macromolecules in cells (**Figure 3–83**), deciphering their full functional meaning may well keep scientists busy for centuries.

## Summary

*Proteins can form enormously sophisticated chemical devices, whose functions largely depend on the detailed chemical properties of their surfaces. Binding sites for ligands are formed as surface cavities in which precisely positioned amino acid side chains are brought together by protein folding. In this way, normally unreactive amino acid side chains can be activated to make and break covalent bonds. Enzymes are catalytic proteins that greatly speed up reaction rates by binding the high-energy transition states for a specific reaction path; they also perform acid catalysis and base catalysis simultaneously. The rates of enzyme reactions are often so fast that they are limited only by diffusion; rates can be further increased if enzymes that act sequentially on a substrate are joined into a single multienzyme complex, or if the enzymes and their substrates are confined to the same compartment of the cell.*

*Proteins reversibly change their shape when ligands bind to their surface. The allosteric changes in protein conformation produced by one ligand affect the binding of a second ligand, and this linkage between two ligand-binding sites provides a crucial mechanism for regulating cell processes. Metabolic pathways, for example, are controlled by feedback regulation: some small molecules inhibit and other small molecules activate enzymes early in a pathway. Enzymes controlled in this way generally form symmetric assemblies, allowing cooperative conformational changes to create a steep response to changes in the concentrations of the ligands that regulate them.*

*The expenditure of chemical energy can drive unidirectional changes in protein shape. By coupling allosteric shape changes to ATP hydrolysis, for example, proteins can do useful work, such as generating a mechanical force or moving for long distances in a single direction. The three-dimensional structures of proteins, determined by x-ray crystallography, have revealed how a small local change caused by nucleoside triphosphate hydrolysis is amplified to create major changes elsewhere in the protein. By such means, these proteins can serve as input–output devices that transmit information, as assembly factors, as motors, or as membrane-bound pumps. Highly efficient protein machines are formed by incorporating many different protein molecules into larger assemblies that coordinate the allosteric movements of the individual components. Such machines are now known to perform many of the most important reactions in cells.*

*Proteins are subjected to many reversible post-translational modifications, such as the covalent addition of a phosphate or an acetyl group to a specific amino acid side chain. The addition of these modifying groups is used to regulate the activity of a protein, changing its conformation, its binding to other proteins and its location inside the cell. A typical protein in a cell will interact with more than five different partners. Using the new technologies of proteomics, biologists can analyze thousands of proteins in one set of experiments. One important result is the production of detailed protein interaction maps, which aim at describing all of the binding interactions between the thousands of distinct proteins in a cell.*

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 72 of 240

# PROBLEMS

**Which statements are true? Explain why or why not.**

**3–1** Each strand in a β sheet is a helix with two amino acids per turn.

**3–2** Loops of polypeptide that protrude from the surface of a protein often form the binding sites for other molecules.

**3–3** An enzyme reaches a maximum rate at high substrate concentration because it has a fixed number of active sites where substrate binds.

**3–4** Higher concentrations of enzyme give rise to a higher turnover number.

**3–5** Enzymes such as aspartate transcarbamoylase that undergo cooperative allosteric transitions invariably contain multiple identical subunits.

**3–6** Continual addition and removal of phosphates by protein kinases and protein phosphatases is wasteful of energy—since their combined action consumes ATP—but it is a necessary consequence of effective regulation by phosphorylation.

**Discuss the following problems.**

**3–7** Consider the following statement. "To produce one molecule of each possible kind of polypeptide chain, 300 amino acids in length, would require more atoms than exist in the universe." Given the size of the universe, do you suppose this statement could possibly be correct? Since counting atoms is a tricky business, consider the problem from the standpoint of mass. The mass of the observable universe is estimated to be about $10^{80}$ grams, give or take an order of magnitude or so. Assuming that the average mass of an amino acid is 110 daltons, what would be the mass of one molecule of each possible kind of polypeptide chain 300 amino acids in length? Is this greater than the mass of the universe?

**3–8** A common strategy for identifying distantly related proteins is to search the database using a short signature sequence indicative of the particular protein function. Why is it better to search with a short sequence than with a long sequence? Do you not have more chances for a 'hit' in the database with a long sequence?

**3–9** The so-called kelch motif consists of a four-stranded β sheet, which forms what is known as a β propeller. It is usually found to be repeated four to seven times, forming a kelch repeat domain in a multidomain protein. One such kelch repeat domain is shown in **Figure Q3–1**. Would you classify this domain as an 'in-line' or 'plug-in' type domain?

**3–10** Titin, which has a molecular weight of $3 \times 10^6$ daltons, is the largest polypeptide yet described. Titin molecules extend from muscle thick filaments to the Z disc; they are thought to act as springs to keep the thick filaments centered in the sarcomere. Titin is composed of a large number of repeated immunoglobulin (Ig) sequences of 89 amino acids, each of which is folded into a domain about 4 nm in length (**Figure Q3–2**A).

You suspect that the springlike behavior of titin is caused by the sequential unfolding (and refolding) of individual Ig domains. You test this hypothesis using the atomic force microscope, which allows you to pick up one end of a protein molecule and pull with an accurately measured force. For a fragment of titin containing seven repeats of the Ig domain, this experiment gives the sawtooth force-versus-extension curve shown in Figure Q3–2B. When the experiment is repeated in a solution of 8 M urea (a protein denaturant), the peaks disappear and the measured extension becomes much longer for a given force. If the experiment is repeated after the protein has been cross-linked by treatment with glutaraldehyde, once again the peaks disappear but the extension becomes much smaller for a given force.

**A.** Are the data consistent with your hypothesis that titin's springlike behavior is due to the sequential unfolding of individual Ig domains? Explain your reasoning.

**B.** Is the extension for each putative domain-unfolding event the magnitude you would expect? (In an extended polypeptide chain, amino acids are spaced at intervals of 0.34 nm.)

**C.** Why is each successive peak in Figure Q3–2B a little higher than the one before?

**D.** Why does the force collapse so abruptly after each peak?

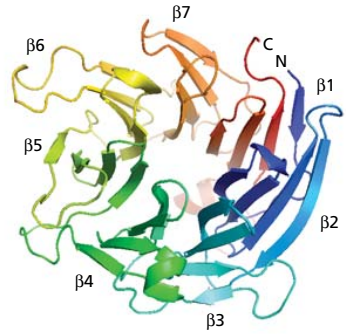



**Figure Q3–1** The kelch repeat domain of galactose oxidase from *D. dendroides* (Problem 3–9). The seven individual β propellers are indicated. The N- and C-termini are indicated by N and C.

**3–11** It is often said that protein complexes are made from subunits (that is, individually synthesized proteins) rather than as one long protein because the former is more likely to give a correct final structure.

**A.** Assuming that the protein synthesis machinery incorporates one incorrect amino acid for each 10,000 it inserts,

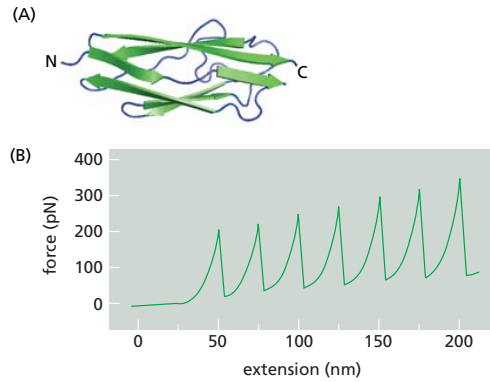



**Figure Q3–2** Springlike behavior of titin (Problem 3–10). (A) The structure of an individual Ig domain. (B) Force in piconewtons versus extension in nanometers obtained by atomic force microscopy.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 73 of 240

calculate the fraction of bacterial ribosomes that would be assembled correctly if the proteins were synthesized as one large protein versus built from individual proteins? For the sake of calculation assume that the ribosome is composed of 50 proteins, each 200 amino acids in length, and that the subunits—correct and incorrect—are assembled with equal likelihood into the complete ribosome. [The probability that a polypeptide will be made correctly, $P_C$, equals the fraction correct for each operation, $f_C$, raised to a power equal to the number of operations, $n$: $P_C = (f_C)^n$. For an error rate of 1/10,000, $f_C = 0.9999$.]

**B.** Is the assumption that correct and incorrect subunits assemble equally well likely to be true? Why or why not? How would a change in that assumption affect the calculation in part A?

**3–12** Rous sarcoma virus (RSV) carries an oncogene called *Src*, which encodes a continuously active protein tyrosine kinase that leads to unchecked cell proliferation. Normally, Src carries an attached fatty acid (myristoylate) group that allows it to bind to the cytoplasmic side of the plasma membrane. A mutant version of Src that does not allow attachment of myristoylate does not bind to the membrane. Infection of cells with RSV encoding either the normal or the mutant form of Src leads to the same high level of protein tyrosine kinase activity, but the mutant Src does not cause cell proliferation.

**A.** Assuming that the normal Src is all bound to the plasma membrane and that the mutant Src is distributed throughout the cytoplasm, calculate their relative concentrations in the neighborhood of the plasma membrane. For the purposes of this calculation, assume that the cell is a sphere with a radius of 10 μm and that the mutant Src is distributed throughout, whereas the normal Src is confined to a 4-nm-thick layer immediately beneath the membrane. [For this problem, assume that the membrane has no thickness. The volume of a sphere is $(4/3)\pi r^3$.]

**B.** The target (X) for phosphorylation by *Src* resides in the membrane. Explain why the mutant Src does not cause cell proliferation.

**3–13** An antibody binds to another protein with an equilibrium constant, $K$, of $5 \times 10^9$ $M^{-1}$. When it binds to a second, related protein, it forms three fewer hydrogen bonds, reducing its binding affinity by 2.8 kcal/mole. What is the $K$ for its binding to the second protein? (Free-energy change is related to the equilibrium constant by the equation $\Delta G° = -2.3\,RT \log K$, where $R$ is $1.98 \times 10^{-3}$ kcal/(mole K) and $T$ is 310 K.)

**3–14** The protein SmpB binds to a special species of tRNA, tmRNA, to eliminate the incomplete proteins made from truncated mRNAs in bacteria. If the binding of SmpB to tmRNA is plotted as fraction tmRNA bound versus SmpB concentration, one obtains a symmetrical S-shaped curve as shown in **Figure Q3–3**. This curve is a visual display of a very useful relationship between $K_d$ and concentration, which has broad applicability. The general expression for fraction of ligand bound is derived from the equation for $K_d$ ($K_d$ = [Pr][L]/[Pr–L]) by substituting ($[L]_{TOT} - [L]$) for [Pr–L] and rearranging. Because the total concentration of ligand ($[L]_{TOT}$) is equal to the free ligand ([L]) plus bound ligand ([Pr–L]),

$$\text{fraction bound} = [L]/[L]_{TOT} = [Pr]/([Pr] + K_d)$$

For SmpB and tmRNA, the fraction bound = $[tmRNA]/[tmRNA]_{TOT} = [SmpB]/([SmpB] + K_d)$. Using this relationship, calculate the fraction of tmRNA bound for SmpB concentrations equal to $10^4K_d$, $10^3K_d$, $10^2K_d$, $10^1K_d$, $K_d$, $10^{-1}K_d$, $10^{-2}K_d$, $10^{-3}K_d$, and $10^{-4}K_d$.

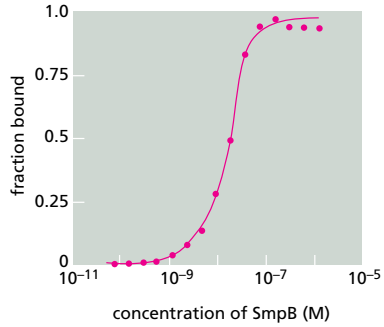



**Figure Q3–3** Fraction of tmRNA bound versus SmpB concentration (Problem 3–14).

**3–15** Many enzymes obey simple Michaelis–Menten kinetics, which are summarized by the equation

$$\text{rate} = V_{max}\,[S]/([S] + K_m)$$

where $V_{max}$ = maximum velocity, [S] = concentration of substrate, and $K_m$ = the Michaelis constant.

It is instructive to plug a few values of [S] into the equation to see how rate is affected. What are the rates for [S] equal to zero, equal to $K_m$, and equal to infinite concentration?

**3–16** The enzyme hexokinase adds a phosphate to D-glucose but ignores its mirror image, L-glucose. Suppose that you were able to synthesize hexokinase entirely from D-amino acids, which are the mirror image of the normal L-amino acids.

**A.** Assuming that the 'D' enzyme would fold to a stable conformation, what relationship would you expect it to bear to the normal 'L' enzyme?

**B.** Do you suppose the 'D' enzyme would add a phosphate to L-glucose, and ignore D-glucose?

**3–17** How do you suppose that a molecule of hemoglobin is able to bind oxygen efficiently in the lungs, and yet release it efficiently in the tissues?

**3–18** Synthesis of the purine nucleotides AMP and GMP proceeds by a branched pathway starting with ribose 5-phosphate (R5P), as shown schematically in **Figure Q3–4**. Using the principles of feedback inhibition, propose a regulatory strategy for this pathway that ensures an adequate supply of both AMP and GMP and minimizes the buildup of the intermediates (*A*–*I*) when supplies of AMP and GMP are adequate.

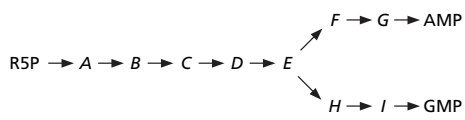



**Figure Q3–4** Schematic diagram of the metabolic pathway for synthesis of AMP and GMP from R5P (Problem 3–18).

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 74 of 240

# REFERENCES


### General

Berg JM, Tymoczko JL & Stryer L (2006) Biochemistry, 6th ed. New York: WH Freeman.
Branden C & Tooze J (1999) Introduction to Protein Structure, 2nd ed. New York: Garland Science.
Dickerson, RE (2005) Present at the Flood: How Structural Molecular Biology Came About. Sunderland, MA: Sinauer
Kyte J (2006) Structure in Protein Chemistry. New York: Routledge.
Petsko GA & Ringe D (2004) Protein Structure and Function. London: New Science Press.
Perutz M (1992) Protein Structure: New Approaches to Disease and Therapy. New York: WH Freeman.

### The Shape and Structure of Proteins

Anfinsen CB (1973) Principles that govern the folding of protein chains. *Science* 181:223–230.
Bray D (2005) Flexible peptides and cytoplasmic gels. *Genome Biol* 6:106–109.
Burkhard P, Stetefeld J & Strelkov SV (2001) Coiled coils: a highly versatile protein folding motif. *Trends Cell Biol* 11:82–88.
Caspar DLD & Klug A (1962) Physical principles in the construction of regular viruses. *Cold Spring Harb Symp Quant Biol* 27:1–24.
Doolittle RF (1995) The multiplicity of domains in proteins. *Annu Rev Biochem* 64:287–314.
Eisenberg D (2003) The discovery of the alpha-helix and beta-sheet, the principle structural features of proteins. *Proc Natl Acad Sci USA* 100:11207–11210.
Fraenkel-Conrat H & Williams RC (1955) Reconstitution of active tobacco mosaic virus from its inactive protein and nucleic acid components. *Proc Natl Acad Sci USA* 41:690–698.
Goodsell DS & Olson AJ (2000) Structural symmetry and protein function. *Annu Rev Biophys Biomol Struct* 29:105–153.
Harrison SC (1992) Viruses. *Curr Opin Struct Biol* 2:293–299.
Harrison SC (2004) Whither structural biology? *Nature Struct Mol Biol* 11:12–15.
Hudder A, Nathanson L & Deutscher MP (2003) Organization of mammalian cytoplasm. *Mol Cell Biol* 23:9318–9326.
International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860–921.
Meiler J & Baker D (2003) Coupled prediction of protein secondary and tertiary structure. *Proc Natl Acad Sci USA* 100:12105–12110.
Nomura M (1973) Assembly of bacterial ribosomes. *Science* 179: 864–873.
Orengo CA & Thornton JM (2005) Protein families and their evolution—a structural perspective. *Annu Rev Biochem* 74:867–900.
Pauling L & Corey RB (1951) Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets. *Proc Natl Acad Sci USA* 37:729–740.
Pauling L, Corey RB & Branson HR (1951) The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci USA* 37:205–211.
Ponting CP, Schultz J, Copley RR et al. (2000) Evolution of domain families. *Adv Protein Chem* 54:185–244.
Trinick J (1992) Understanding the functions of titin and nebulin. *FEBS Lett* 307:44–48.
Vogel C, Bashton M, Kerrison ND et al (2004) Structure, function and evolution of multidomain proteins. *Curr Opin Struct Biol* 14:208–216.
Zhang C & Kim SH (2003) Overview of structural genomics: from structure to function. *Curr Opin Chem Biol* 7:28–32.

### Protein Function

Alberts B (1998) The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell* 92:291–294.
Benkovic SJ (1992) Catalytic antibodies. *Annu Rev Biochem* 61:29–54.
Berg OG & von Hippel PH (1985) Diffusion-controlled macromolecular interactions. *Annu Rev Biophys Biophys Chem* 14:131–160.
Bhattacharyya RP, Remenyi A, Yeh BJ & Lim WA (2006) Domains, motifs, and scaffolds: The role of modular interactions in the evolution and wiring of cell signaling circuits. *Annu Rev Biochem* 75:655–680.
Bourne HR (1995) GTPases: a family of molecular switches and clocks. *Philos Trans R Soc Lond B* 349:283–289.
Braden BC & Poljak RJ (1995) Structural features of the reactions between antibodies and protein antigens. *FASEB J* 9:9–16.
Dickerson RE & Geis I (1983) Hemoglobin: Structure, Function, Evolution and Pathology. Menlo Park, CA: Benjamin Cummings.
Dressler D & Potter H (1991) Discovering Enzymes. New York: Scientific American Library.
Eisenberg D, Marcotte, EM, Xenarios, I & Yeates TO (2000) Protein function in the post-genomic era. *Nature* 405:823–826.
Fersht AR (1999) Structure and Mechanisms in Protein Science: A Guide to Enzyme Catalysis. New York: WH Freeman.
Johnson, LN & Lewis RJ (2001) Structural basis for control by phosphorylation. *Chem Rev* 101:2209–2242.
Kantrowitz ER & Lipscomb WN (1988) *Escherichia coli* aspartate transcarbamoylase: the relation between structure and function. *Science* 241:669–674.
Khosla C & Harbury PB (2001) Modular enzymes. *Nature* 409:247–252.
Kim E & Sheng M (2004) PDZ domain proteins of synapses. *Nature Rev Neurosci* 5:771–781.
Koshland DE, Jr (1984) Control of enzyme activity and metabolic pathways. *Trends Biochem Sci* 9:155–159.
Kraut DA, Carroll KS & Herschlag D (2003) Challenges in enzyme mechanism and energetics. *Annu Rev Biochem* 72:517–571.
Krogan NJ, Cagney G, Yu H et al (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisisae*. *Nature* 440:637–643.
Lichtarge O, Bourne HR & Cohen FE (1996) An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol* 257:342–358.
Marcotte EM, Pellegrini M, Ng HL et al (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science* 285:751–753.
Monod J, Changeux JP & Jacob F (1963) Allosteric proteins and cellular control systems. *J Mol Biol* 6:306–329.
Pawson T & Nash P (2003) Assembly of regulatory systems through protein interaction domains. *Science* 300:445–452.
Pavletich NP (1999) Mechanisms of cyclin-dependent kinase regulation: structures of Cdks, their cyclin activators, and Cip and INK4 inhibitors. *J Mol Biol* 287:821–828.
Pellicena P & Kuriyan J (2006) Protein-protein interactions in the allosteric regulation of protein kinases. *Curr Opin Struct Biol* 16:702–709.
Perutz M (1990) Mechanisms of Cooperativity and Allosteric Regulation in Proteins. Cambridge: Cambridge University Press.
Raushel F, Thoden, JB & Holden HM (2003) Enzymes with molecular tunnels. *Acc Chem Res* 36:539–548.
Radzicka A & Wolfenden R (1995) A proficient enzyme. *Science* 267:90–93.
Sato TK, Overduin M & Emr S (2001) Location, location, location: Membrane targeting directed by PX domains. *Science* 294:1881–1885.
Schramm VL (1998) Enzymatic transition states and transition state analog design. *Annu Rev Biochem* 67:693–720.
Schultz PG & Lerner RA (1995) From molecular diversity to catalysis: lessons from the immune system. *Science* 269:1835–1842.
Vale RD & Milligan RA (2000) The way things move: looking under the hood of molecular motor proteins. *Science* 288:88–95.
Vocadlo DJ, Davies GJ, Laine R & Withers SG (2001) Catalysis by hen egg-white lysozyme proceeds via a covalent intermediate. *Nature* 412:835–838.
Walsh C (2001) Enabling the chemistry of life. *Nature* 409:226–231.
Yang XJ (2005) Multisite protein modification and intramolecular signaling. *Oncogene* 24:1653–1662.
Zhu H, Bilgin M & Snyder M (2003) Proteomics. *Annu Rev Biochem* 72:783–812.


LCY Biotechnology Holding, Inc.
Ex. 1008
Page 75 of 240

# How Cells Read the Genome: From DNA to Protein

6

Only when the structure of DNA was discovered in the early 1950s did it become clear how the hereditary information in cells is encoded in DNA's sequence of nucleotides. The progress since then has been astounding. Within fifty years we knew the complete genome sequences for many organisms, including humans. We therefore know the maximum amount of information that is required to produce a complex organism like ourselves. The limits on the hereditary information needed for life constrain the biochemical and structural features of cells and make it clear that biology is not infinitely complex.

In this chapter, we explain how cells decode and use the information in their genomes. Much has been learned about how the genetic instructions written in an alphabet of just four "letters"—the four different nucleotides in DNA—direct the formation of a bacterium, a fruit fly, or a human. Nevertheless, we still have a great deal to discover about how the information stored in an organism's genome produces even the simplest unicellular bacterium with 500 genes, let alone how it directs the development of a human with approximately 25,000 genes. An enormous amount of ignorance remains; many fascinating challenges therefore await the next generation of cell biologists.

The problems that cells face in decoding genomes can be appreciated by considering a small portion of the genome of the fruit fly *Drosophila melanogaster* (**Figure 6–1**). Much of the DNA-encoded information present in this and other genomes specifies the linear order—the sequence—of amino acids for every protein the organism makes. As described in Chapter 3, the amino acid sequence in turn dictates how each protein folds to give a molecule with a distinctive shape and chemistry. When a cell makes a particular protein, it must decode accurately the corresponding region of the genome. Additional information encoded in the DNA of the genome specifies exactly when in the life of an organism and in which cell types each gene is to be expressed into protein. Since proteins are the main constituents of cells, the decoding of the genome determines not only the size, shape, biochemical properties, and behavior of cells, but also the distinctive features of each species on Earth.

One might have predicted that the information present in genomes would be arranged in an orderly fashion, resembling a dictionary or a telephone directory. Although the genomes of some bacteria seem fairly well organized, the genomes of most multicellular organisms, such as our *Drosophila* example, are surprisingly disorderly. Small bits of coding DNA (that is, DNA that codes for protein) are interspersed with large blocks of seemingly meaningless DNA. Some sections of the genome contain many genes and others lack genes altogether. Proteins that work closely with one another in the cell often have their genes located on different chromosomes, and adjacent genes typically encode proteins that have little to do with each other in the cell. Decoding genomes is therefore no simple matter. Even with the aid of powerful computers, it is still difficult for researchers to locate definitively the beginning and end of genes in the DNA sequences of complex genomes, much less to predict when each gene is expressed in the life of the organism. Although the DNA sequence of the human genome is known, it will probably take at least a decade to identify every gene and determine the precise amino acid sequence of the protein it produces. Yet the cells in our body do this thousands of times a second.

## In This Chapter

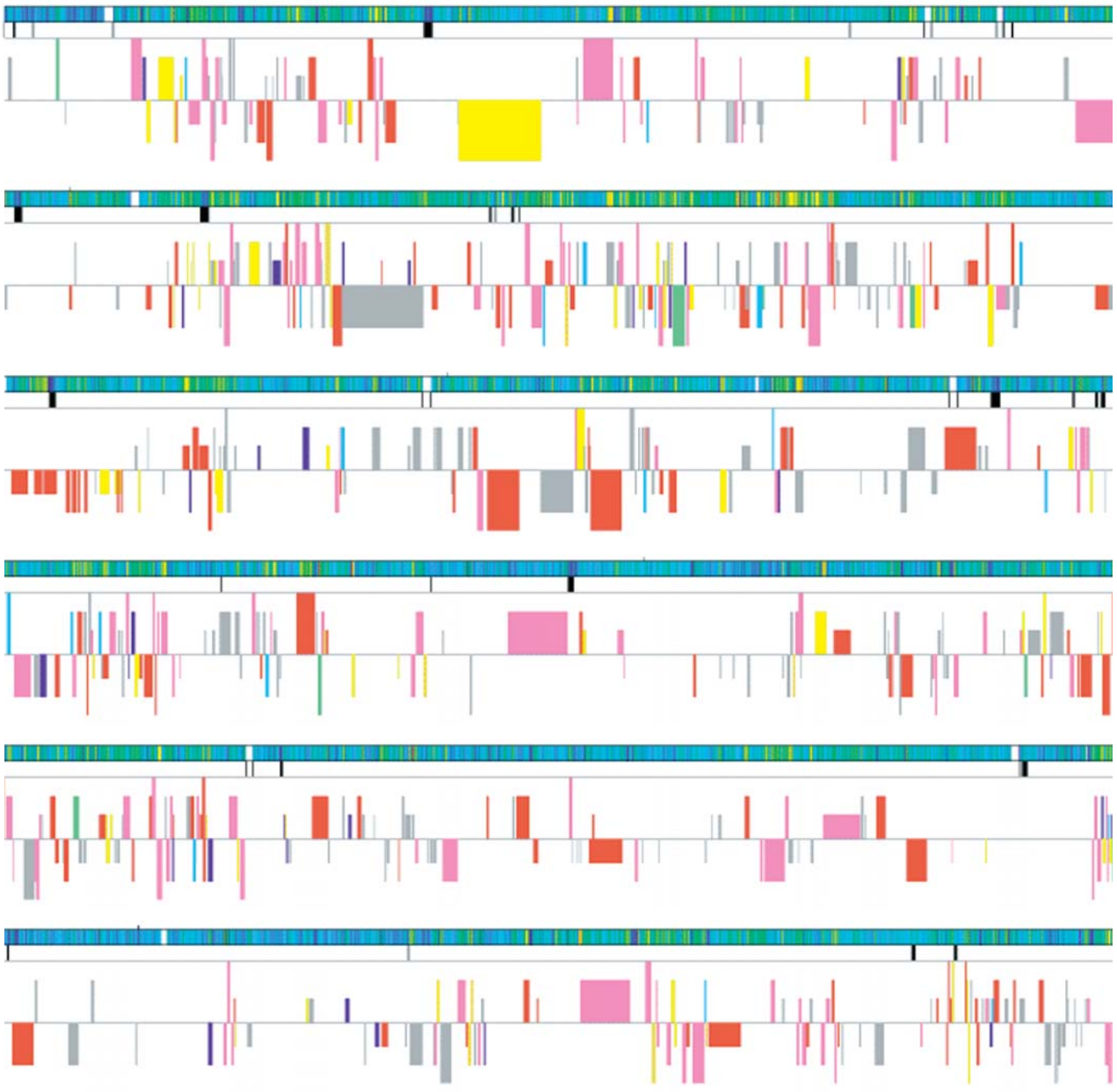
FROM DNA TO RNA 331

FROM RNA TO PROTEIN 366

THE RNA WORLD AND THE ORIGINS OF LIFE 400



LCY Biotechnology Holding, Inc.
Ex. 1008
Page 76 of 240



KEY:

13 or more
1 to 12
none
length of bar indicates number of corresponding cDNAs identified in databases

%GC content
25 65
transposable elements
known and predicted genes identified on top strand of DNA
known and predicted genes identified on bottom strand of DNA

100,000 nucleotide pairs

color code for sequence similarity of genes identified

MWY
WY
MW
W
MY
Y
M
no similarity to MWY

M = mammalian
W = *C. elegans*
Y = *S. cerevisiae*

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 77 of 240

**Figure 6–1** ***(opposite page)*** **Schematic depiction of a portion of chromosome 2 from the genome of the fruit fly *Drosophila melanogaster*.** This figure represents approximately 3% of the total *Drosophila* genome, arranged as six contiguous segments. As summarized in the key, the symbolic representations are: *black vertical lines* of various thicknesses: locations of transposable elements, with thicker bars indicating clusters of elements; *colored boxes*: genes (both known and predicted) coded on one strand of DNA (boxes *above* the midline) and genes coded on the other strand (boxes *below* the midline). The length of each gene box includes both its exons (protein-coding DNA) and its introns (noncoding DNA) (see Figure 4–15); its height is proportional to the number of known cDNAs that match the gene. (As described in Chapter 8, cDNAs are DNA copies of mRNA molecules, and large collections of the nucleotide sequences of cDNAs have been deposited in a variety of databases, the more matches, the higher the confidence that the predicted gene is transcribed into RNA and is thus a genuine gene.) The color of each gene box indicates whether a closely related gene is known to occur in other organisms. For example, MWY means the gene has close relatives in mammals, in the nematode worm *Caenorhabditis elegans*, and in the yeast *Saccharomyces cerevisiae*. MW indicates the gene has close relatives in mammals and the worm but not in yeast. The *rainbow-colored bar* indicates percent G–C base pairs; across many different genomes, this percentage shows a striking regional variation, whose origin and significance are uncertain. (From M.D. Adams et al., *Science* 287:2185–2195, 2000. With permission from AAAS.)

The DNA in genomes does not direct protein synthesis itself, but instead uses RNA as an intermediary. When the cell needs a particular protein, the nucleotide sequence of the appropriate portion of the immensely long DNA molecule in a chromosome is first copied into RNA (a process called *transcription)*. It is these RNA copies of segments of the DNA that are used directly as templates to direct the synthesis of the protein (a process called *translation)*. The flow of genetic information in cells is therefore from DNA to RNA to protein (**Figure 6–2**). All cells, from bacteria to humans, express their genetic information in this way—a principle so fundamental that it is termed the *central dogma* of molecular biology.

Despite the universality of the central dogma, there are important variations in the way in which information flows from DNA to protein. Principal among these is that RNA transcripts in eucaryotic cells are subject to a series of processing steps in the nucleus, including *RNA splicing*, before they are permitted to exit from the nucleus and be translated into protein. These processing steps can critically change the "meaning" of an RNA molecule and are therefore crucial for understanding how eucaryotic cells read their genomes. Finally, although we focus on the production of the proteins encoded by the genome in this chapter, we see that for many genes RNA is the final product. Like proteins, many of these RNAs fold into precise three-dimensional structures that have structural, catalytic, and regulatory roles in the cell.

We begin this chapter with the first step in decoding a genome: the process of transcription by which an RNA molecule is produced from the DNA of a gene. We then follow the fate of this RNA molecule through the cell, finishing when a correctly folded protein molecule has been formed. At the end of the chapter, we consider how the present quite complex scheme of information storage, transcription, and translation might have arisen from simpler systems in the earliest stages of cell evolution.

## FROM DNA TO RNA

Transcription and translation are the means by which cells read out, or express, the genetic instructions in their genes. Because many identical RNA copies can be made from the same gene, and each RNA molecule can direct the synthesis of many identical protein molecules, cells can synthesize a large amount of protein rapidly when necessary. But each gene can also be transcribed and translated with a different efficiency, allowing the cell to make vast quantities of some proteins and tiny quantities of others (**Figure 6–3**). Moreover, as we see in the next chapter, a cell can change (or regulate) the expression of each of its genes according to the needs of the moment—most commonly by controlling the production of its RNA.

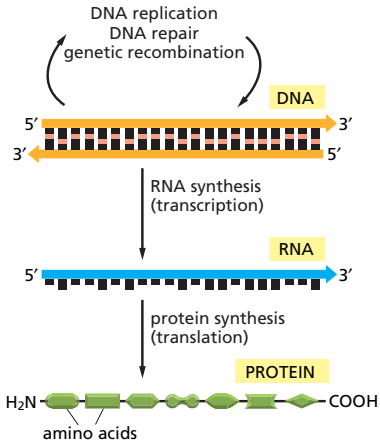



**Figure 6–2 The pathway from DNA to protein.** The flow of genetic information from DNA to RNA (transcription) and from RNA to protein (translation) occurs in all living cells.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 78 of 240

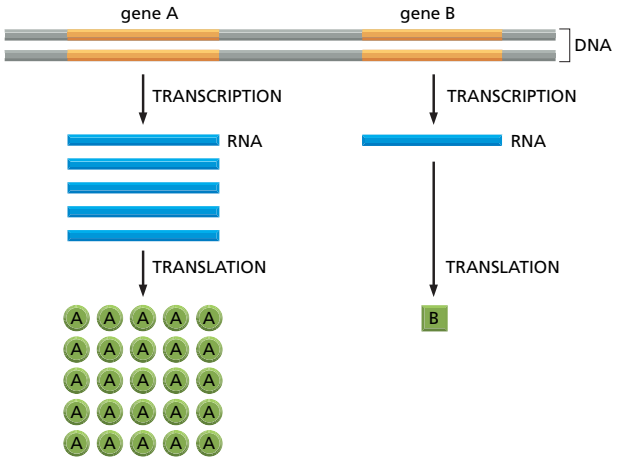


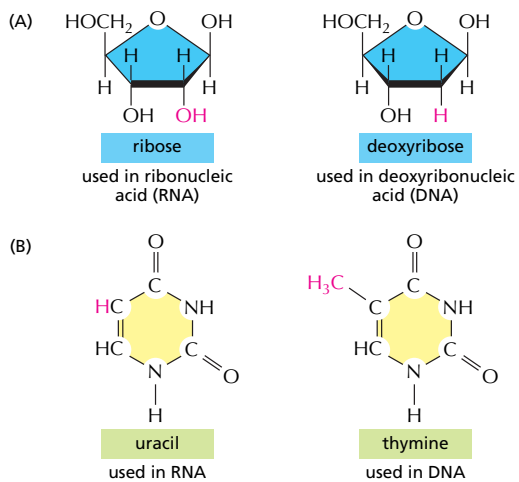

**Figure 6–3 Genes can be expressed with different efficiencies.** In this example, gene A is transcribed and translated much more efficiently than gene B. This allows the amount of protein A in the cell to be much greater than that of protein B.

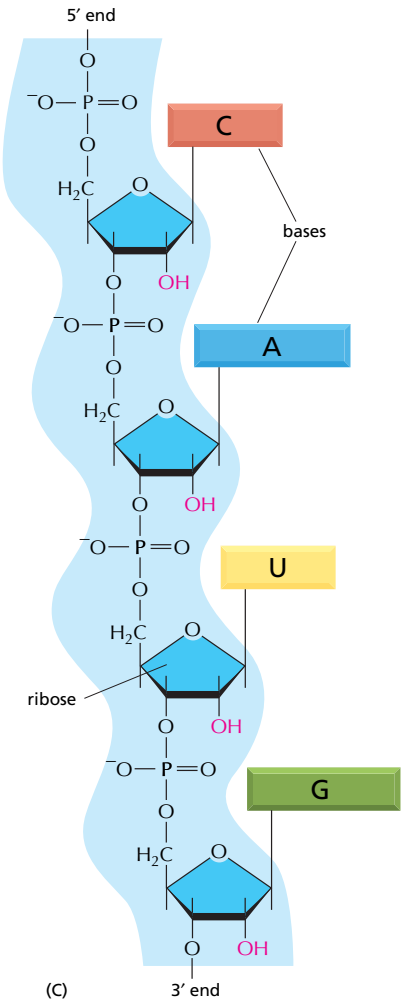
## Portions of DNA Sequence Are Transcribed into RNA

The first step a cell takes in reading out a needed part of its genetic instructions is to copy a particular portion of its DNA nucleotide sequence—a gene—into an RNA nucleotide sequence. The information in RNA, although copied into another chemical form, is still written in essentially the same language as it is in DNA—the language of a nucleotide sequence. Hence the name **transcription**.

Like DNA, RNA is a linear polymer made of four different types of nucleotide subunits linked together by phosphodiester bonds (**Figure 6–4**). It differs from DNA chemically in two respects: (1) the nucleotides in RNA are *ribonucleotides*—that is, they contain the sugar ribose (hence the name *ribo*nucleic acid) rather than deoxyribose; (2) although, like DNA, RNA contains the bases adenine (A), guanine (G), and cytosine (C), it contains the base uracil (U) instead of the thymine (T) in DNA. Since U, like T, can base-pair by hydrogen-bonding with A (**Figure 6–5**), the complementary base-pairing properties described for DNA in Chapters 4 and 5 apply also to RNA (in RNA, G pairs with C, and A pairs with U). We also find other types of base pairs in RNA: for example, G occasionally pairs with U.

**Figure 6–4 The chemical structure of RNA.** (A) RNA contains the sugar ribose, which differs from deoxyribose, the sugar used in DNA, by the presence of an additional –OH group. (B) RNA contains the base uracil, which differs from thymine, the equivalent base in DNA, by the absence of a $-CH_3$ group. (C) A short length of RNA. The phosphodiester chemical linkage between nucleotides in RNA is the same as that in DNA.




LCY Biotechnology Holding, Inc.
Ex. 1008
Page 79 of 240

Although these chemical differences are slight, DNA and RNA differ quite dramatically in overall structure. Whereas DNA always occurs in cells as a double-stranded helix, RNA is single-stranded. An RNA chain can therefore fold up into a particular shape, just as a polypeptide chain folds up to form the final shape of a protein (**Figure 6–6**). As we see later in this chapter, the ability to fold into complex three-dimensional shapes allows some RNA molecules to have precise structural and catalytic functions.

## Transcription Produces RNA Complementary to One Strand of DNA

The RNA in a cell is made by DNA transcription, a process that has certain similarities to the process of DNA replication discussed in Chapter 5. Transcription begins with the opening and unwinding of a small portion of the DNA double helix to expose the bases on each DNA strand. One of the two strands of the DNA double helix then acts as a template for the synthesis of an RNA molecule. As in DNA replication, the nucleotide sequence of the RNA chain is determined by the complementary base-pairing between incoming nucleotides and the DNA template. When a good match is made, the incoming ribonucleotide is covalently linked to the growing RNA chain in an enzymatically catalyzed reaction. The RNA chain produced by transcription—the *transcript*—is therefore elongated one nucleotide at a time, and it has a nucleotide sequence that is exactly complementary to the strand of DNA used as the template (**Figure 6–7**).

Transcription, however, differs from DNA replication in several crucial ways. Unlike a newly formed DNA strand, the RNA strand does not remain hydrogen-bonded to the DNA template strand. Instead, just behind the region where the ribonucleotides are being added, the RNA chain is displaced and the DNA helix re-forms. Thus, the RNA molecules produced by transcription are released from the DNA template as single strands. In addition, because they are copied from only a limited region of the DNA, RNA molecules are much shorter than DNA molecules. A DNA molecule in a human chromosome can be up to 250 million nucleotide-pairs long; in contrast, most RNAs are no more than a few thousand nucleotides long, and many are considerably shorter.

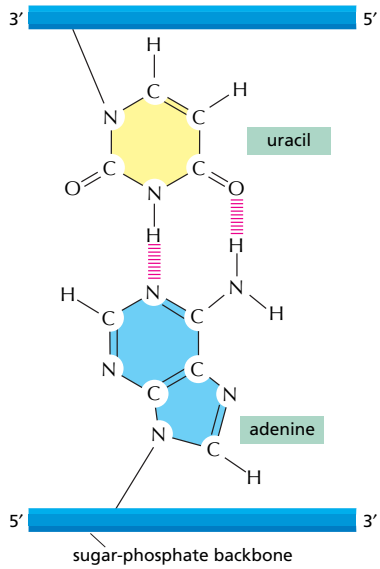



**Figure 6–5 Uracil forms base pairs with adenine.** The absence of a methyl group in U has no effect on base-pairing; thus, U–A base pairs closely resemble T–A base pairs (see Figure 4–4).

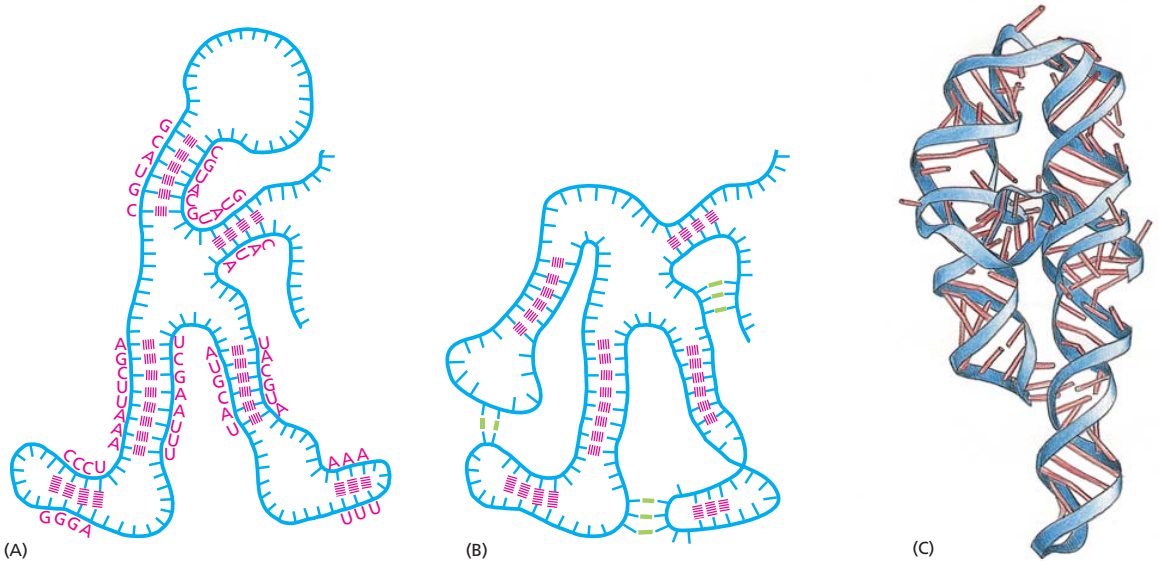



**Figure 6–6 RNA can fold into specific structures.** RNA is largely single-stranded, but it often contains short stretches of nucleotides that can form conventional base pairs with complementary sequences found elsewhere on the same molecule. These interactions, along with additional "nonconventional" base-pair interactions, allow an RNA molecule to fold into a three-dimensional structure that is determined by its sequence of nucleotides. <AATC> (A) Diagram of a folded RNA structure showing only conventional base-pair interactions. (B) Structure with both conventional *(red)* and nonconventional *(green)* base-pair interactions. (C) Structure of an actual RNA, a portion of a group I intron (see Figure 6–36). Each conventional base-pair interaction is indicated by a "rung" in the double helix. Bases in other configurations are indicated by broken rungs.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 80 of 240

The enzymes that perform transcription are called **RNA polymerases**. Like the DNA polymerase that catalyzes DNA replication (discussed in Chapter 5), RNA polymerases catalyze the formation of the phosphodiester bonds that link the nucleotides together to form a linear chain. The RNA polymerase moves stepwise along the DNA, unwinding the DNA helix just ahead of the active site for polymerization to expose a new region of the template strand for complementary base-pairing. In this way, the growing RNA chain is extended by one nucleotide at a time in the 5′-to-3′ direction (**Figure 6–8**). The substrates are nucleoside triphosphates (ATP, CTP, UTP, and GTP); as in DNA replication, the hydrolysis of high-energy bonds provides the energy needed to drive the reaction forward (see Figure 5–4).

The almost immediate release of the RNA strand from the DNA as it is synthesized means that many RNA copies can be made from the same gene in a relatively short time, with the synthesis of additional RNA molecules being started before the first RNA is completed (**Figure 6–9**). When RNA polymerase molecules follow hard on each other's heels in this way, each moving at about 20 nucleotides per second (the speed in eucaryotes), over a thousand transcripts can be synthesized in an hour from a single gene.

Although RNA polymerase catalyzes essentially the same chemical reaction as DNA polymerase, there are some important differences between the activities of the two enzymes. First, and most obviously, RNA polymerase catalyzes the linkage of ribonucleotides, not deoxyribonucleotides. Second, unlike the DNA polymerases involved in DNA replication, RNA polymerases can start an RNA chain without a primer. This difference may exist because transcription need not be as accurate as DNA replication (see Table 5–1, p. 271). Unlike DNA, RNA does not permanently store genetic information in cells. RNA polymerases make about one mistake for every $10^4$ nucleotides copied into RNA (compared with an error rate for direct copying by DNA polymerase of about one in $10^7$ nucleotides), and the consequences of an error in RNA transcription are much less significant than that in DNA replication.

Although RNA polymerases are not nearly as accurate as the DNA polymerases that replicate DNA, they nonetheless have a modest proofreading mechanism. If an incorrect ribonucleotide is added to the growing RNA chain, the polymerase can back up, and the active site of the enzyme can perform an excision reaction that resembles the reverse of the polymerization reaction,

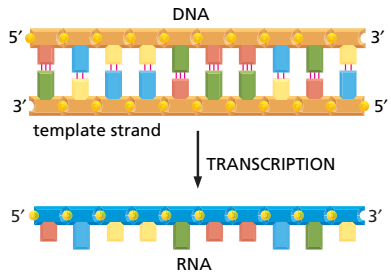



**Figure 6–7 DNA transcription produces a single-stranded RNA molecule that is complementary to one strand of DNA.**

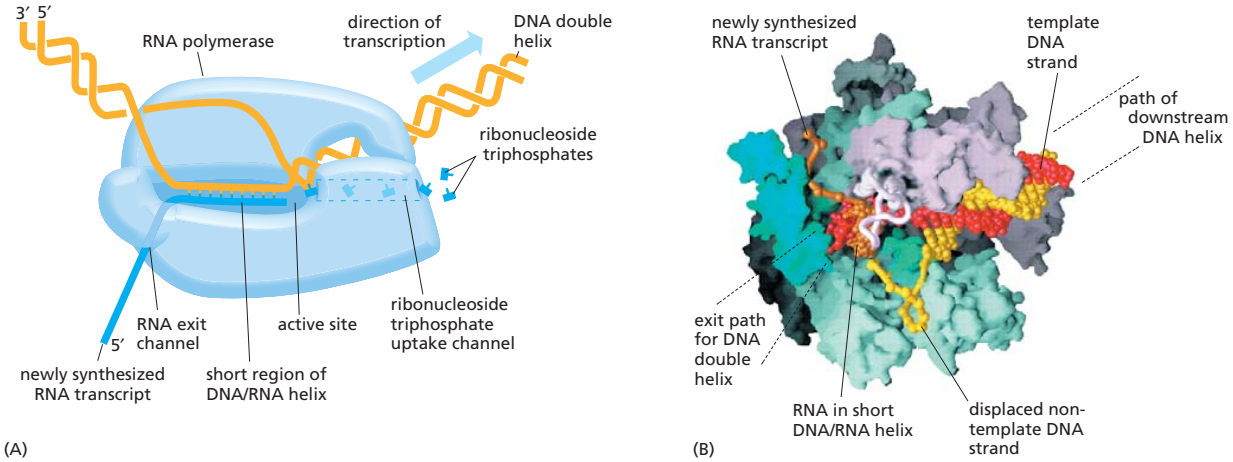



**Figure 6–8 DNA is transcribed by the enzyme RNA polymerase.** (A) The RNA polymerase *(pale blue)* moves stepwise along the DNA, unwinding the DNA helix at its active site. As it progresses, the polymerase adds nucleotides (represented as *small "T" shapes)* one by one to the RNA chain at the polymerization site, using an exposed DNA strand as a template. The RNA transcript is thus a complementary copy of one of the two DNA strands. A short region of DNA/RNA helix (approximately nine nucleotide pairs in length) is therefore formed only transiently, and a "window" of DNA/RNA helix therefore moves along the DNA with the polymerase. The incoming nucleotides are in the form of ribonucleoside triphosphates (ATP, UTP, CTP, and GTP), and the energy stored in their phosphate–phosphate bonds provides the driving force for the polymerization reaction (see Figure 5–4). (B) The structure of a bacterial RNA polymerase, as determined by x-ray crystallography. Four different subunits, indicated by different colors, comprise this RNA polymerase. The DNA strand used as a template is *red*, and the nontemplate strand is *yellow*. (A, adapted from a figure courtesy of Robert Landick; B, courtesy of Seth Darst.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 81 of 240

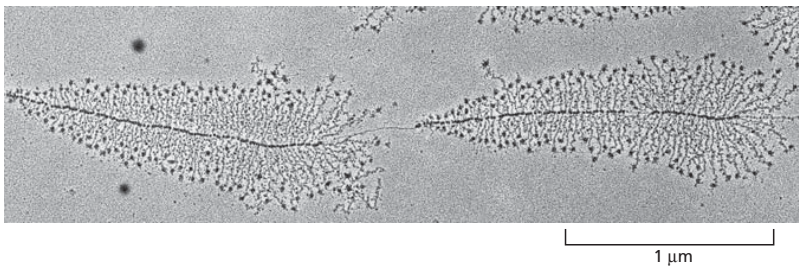



**Figure 6–9 Transcription of two genes as observed under the electron microscope.** The micrograph shows many molecules of RNA polymerase simultaneously transcribing each of two adjacent genes. Molecules of RNA polymerase are visible as a series of dots along the DNA with the newly synthesized transcripts (fine threads) attached to them. The RNA molecules (ribosomal RNAs) shown in this example are not translated into protein but are instead used directly as components of ribosomes, the machines on which translation takes place. The particles at the 5′ end (the free end) of each rRNA transcript are believed to reflect the beginnings of ribosome assembly. From the lengths of the newly synthesized transcripts, it can be deduced that the RNA polymerase molecules are transcribing from left to right. (Courtesy of Ulrich Scheer.)

except that water instead of pyrophosphate is used and a nucleoside monophosphate is released.

Given that DNA and RNA polymerases both carry out template-dependent nucleotide polymerization, it might be expected that the two types of enzymes would be structurally related. However, x-ray crystallographic studies of both types of enzymes reveal that, other than containing a critical $Mg^{2+}$ ion at the catalytic site, they are virtually unrelated to each other; indeed template-dependent nucleotide polymerizing enzymes seem to have arisen independently twice during the early evolution of cells. One lineage led to the modern DNA polymerases and reverse transcriptases discussed in Chapter 5, as well as to a few single-subunit RNA polymerases from viruses. The other lineage formed all of the modern cellular RNA polymerases (**Figure 6–10**), which we discuss in this chapter.

## Cells Produce Several Types of RNA

The majority of genes carried in a cell's DNA specify the amino acid sequence of proteins; the RNA molecules that are copied from these genes (which ultimately direct the synthesis of proteins) are called **messenger RNA** (**mRNA**) molecules. The final product of a minority of genes, however, is the RNA itself. Careful analysis of the complete DNA sequence of the genome of the yeast *S. cerevisiae* has uncovered well over 750 genes (somewhat more than 10% of the total number of yeast genes) that produce RNA as their final product. These RNAs, like proteins, serve as enzymatic and structural components for a wide variety of processes in

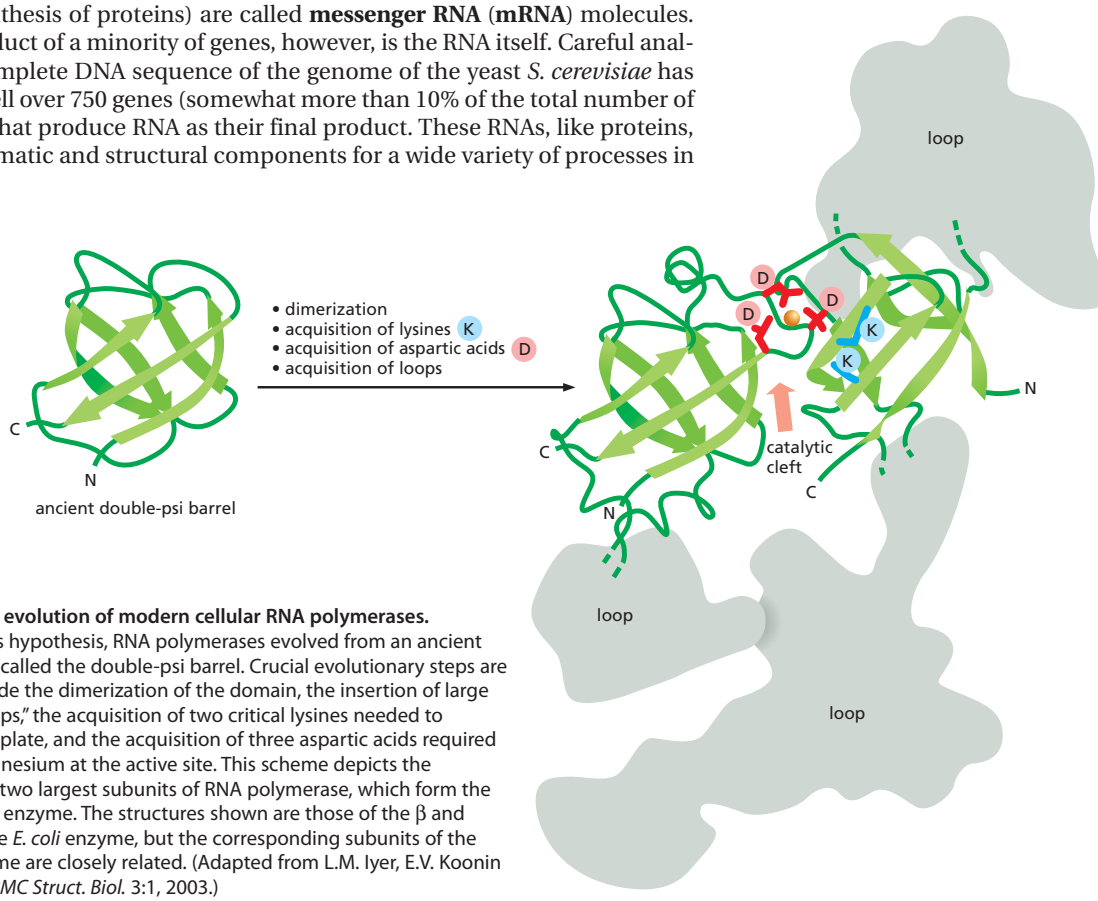



**Figure 6–10 The evolution of modern cellular RNA polymerases.** According to this hypothesis, RNA polymerases evolved from an ancient protein domain, called the double-psi barrel. Crucial evolutionary steps are thought to include the dimerization of the domain, the insertion of large polypeptide "loops," the acquisition of two critical lysines needed to position the template, and the acquisition of three aspartic acids required to chelate a magnesium at the active site. This scheme depicts the evolution of the two largest subunits of RNA polymerase, which form the active site of the enzyme. The structures shown are those of the β and β′ subunits of the *E. coli* enzyme, but the corresponding subunits of the eucaryotic enzyme are closely related. (Adapted from L.M. Iyer, E.V. Koonin and L. Aravind, *BMC Struct. Biol.* 3:1, 2003.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 82 of 240

the cell. In Chapter 5 we encountered one of those RNAs, the template carried by the enzyme telomerase. Although many of these noncoding RNAs are still mysterious, we shall see in this chapter that *small nuclear RNA (snRNA)* molecules direct the splicing of pre-mRNA to form mRNA, that *ribosomal RNA (rRNA)* molecules form the core of ribosomes, and that *transfer RNA (tRNA)* molecules form the adaptors that select amino acids and hold them in place on a ribosome for incorporation into protein. Finally, we shall see in Chapter 7 that *microRNA (miRNA)* molecules and *small interfering RNA (siRNA)* molecules serve as key regulators of eucaryotic gene expression (**Table 6–1**).

Each transcribed segment of DNA is called a *transcription unit*. In eucaryotes, a transcription unit typically carries the information of just one gene, and therefore codes for either a single RNA molecule or a single protein (or group of related proteins if the initial RNA transcript is spliced in more than one way to produce different mRNAs). In bacteria, a set of adjacent genes is often transcribed as a unit; the resulting mRNA molecule therefore carries the information for several distinct proteins.

Overall, RNA makes up a few percent of a cell's dry weight. Most of the RNA in cells is rRNA; mRNA comprises only 3–5% of the total RNA in a typical mammalian cell. The mRNA population is made up of tens of thousands of different species, and there are on average only 10–15 molecules of each species of mRNA present in each cell.

## Signals Encoded in DNA Tell RNA Polymerase Where to Start and Stop

To transcribe a gene accurately, RNA polymerase must recognize where on the genome to start and where to finish. The way in which RNA polymerases perform these tasks differs somewhat between bacteria and eucaryotes. Because the processes in bacteria are simpler, we discuss them first.

The initiation of transcription is an especially important step in gene expression because it is the main point at which the cell regulates which proteins are to be produced and at what rate. The bacterial RNA polymerase core enzyme is a multisubunit complex that sythesizes RNA using a DNA template as a guide. A detachable subunit called *sigma (σ) factor* associates with the core enzyme and assists it in reading the signals in the DNA that tell it where to begin transcribing (**Figure 6–11**). Together, $\sigma$ factor and core enzyme are known as the **RNA polymerase holoenzyme**; this complex adheres only weakly to bacterial DNA when

**Table 6–1 Principal Types of RNAs Produced in Cells**

| TYPE OF RNA | FUNCTION |
|---|---|
| mRNAs | messenger RNAs, code for proteins |
| rRNAs | ribosomal RNAs, form the basic structure of the ribosome and catalyze protein synthesis |
| tRNAs | transfer RNAs, central to protein synthesis as adaptors between mRNA and amino acids |
| snRNAs | small nuclear RNAs, function in a variety of nuclear processes, including the splicing of pre-mRNA |
| snoRNAs | small nucleolar RNAs, used to process and chemically modify rRNAs |
| scaRNAs | small cajal RNAs, used to modify snoRNAs and snRNAs |
| miRNAs | microRNAs, regulate gene expression typically by blocking translation of selective mRNAs |
| siRNAs | small interfering RNAs, turn off gene expression by directing degradation of selective mRNAs and the establishment of compact chromatin structures |
| Other noncoding RNAs | function in diverse cell processes, including telomere synthesis, X-chromosome inactivation, and the transport of proteins into the ER |

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 83 of 240

the two collide, and a holoenzyme typically slides rapidly along the long DNA molecule until it dissociates again. However, when the polymerase holoenzyme slides into a region on the DNA double helix called a **promoter**, a special sequence of nucleotides indicating the starting point for RNA synthesis, the polymerase binds tightly to this DNA. The polymerase holoenzyme, through its σ factor, recognizes the promoter DNA sequence by making specific contacts with the portions of the bases that are exposed on the outside of the helix (step 1 in Figure 6–11).

After the RNA polymerase holoenzyme binds tightly to the promoter DNA in this way, it opens up the double helix to expose a short stretch of nucleotides on each strand (step 2 in Figure 6–11). Unlike a DNA helicase reaction (see Figure 5–14), this limited opening of the helix does not require the energy of ATP hydrolysis. Instead, the polymerase and DNA both undergo reversible structural changes that result in a state more energetically favorable than that of the initial binding. With the DNA unwound, one of the two exposed DNA strands acts as a template for complementary base-pairing with incoming ribonucleotides, two of which are joined together by the polymerase to begin an RNA chain (step 3 in Figure 6–11). After the first ten or so nucleotides of RNA have been synthesized (a relatively inefficient process during which polymerase synthesizes and discards short RNA oligomers), the core enzyme breaks its interactions with the promoter DNA, weakens its interactions with σ factor, and begins to move down the DNA, synthesizing RNA (steps 4 and 5 in Figure 6–11). Chain elongation continues (at a speed of approximately 50 nucleotides/sec for bacterial RNA polymerases) until the enzyme encounters a second signal in the DNA, the **terminator** (described below), where the polymerase halts and releases both the newly made RNA chain and the DNA template (step 7 in Figure 6–11). After the polymerase core enzyme has been released at a terminator, it reassociates with a free σ factor to form a holoenzyme that can begin the process of transcription again.

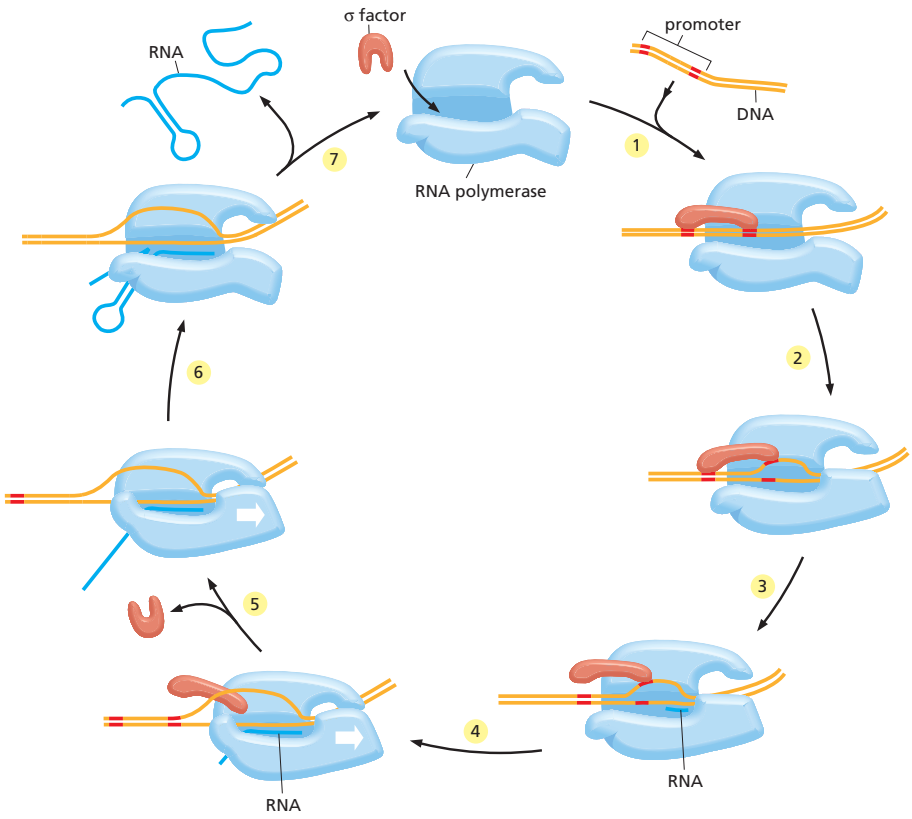
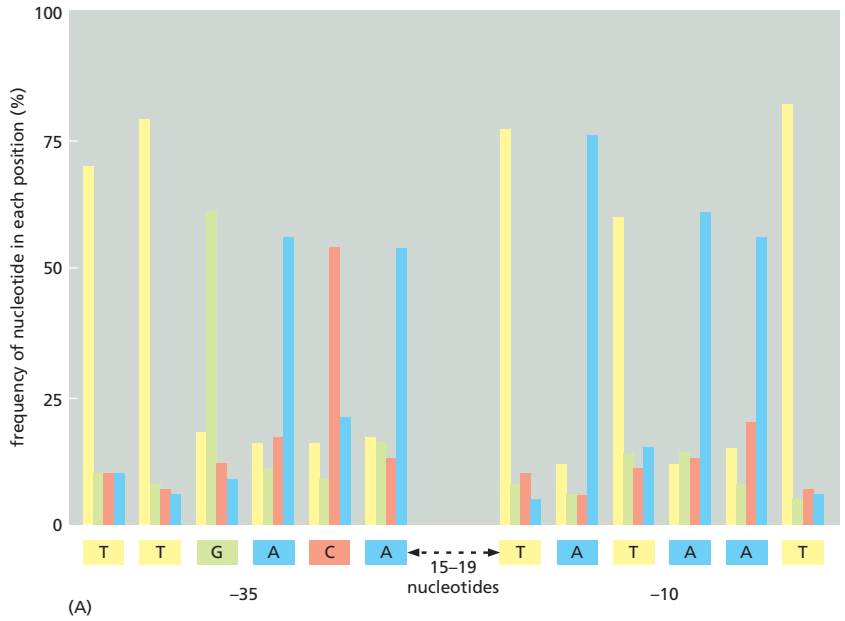



**Figure 6–11 The transcription cycle of bacterial RNA polymerase.** In step 1, the RNA polymerase holoenzyme (polymerase core enzyme plus σ factor) assembles and then locates a promoter (see Figure 6–12). The polymerase unwinds the DNA at the position at which transcription is to begin (step 2) and begins transcribing (step 3). This initial RNA synthesis (sometimes called "abortive initiation") is relatively inefficient. However, once RNA polymerase has managed to synthesize about 10 nucleotides of RNA, it breaks its interactions with the promoter DNA and weakens, and eventually breaks, its interaction with σ. The polymerase now shifts to the elongation mode of RNA synthesis (step 4), moving rightward along the DNA in this diagram. During the elongation mode (step 5), transcription is highly processive, with the polymerase leaving the DNA template and releasing the newly transcribed RNA only when it encounters a termination signal (steps 6 and 7). Termination signals are typically encoded in DNA, and many function by forming an RNA structure that destabilizes the polymerase's hold on the RNA (step 7). In bacteria, all RNA molecules are synthesized by a single type of RNA polymerase and the cycle depicted in the figure therefore applies to the production of mRNAs as well as structural and catalytic RNAs. (Adapted from a figure courtesy of Robert Landick.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 84 of 240



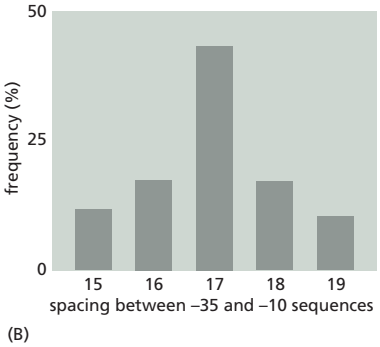




**Figure 6–12 Consensus sequence for the major class of *E. coli* promoters.** (A) The promoters are characterized by two hexameric DNA sequences, the –35 sequence and the –10 sequence named for their approximate location relative to the start point of transcription (designated +1). For convenience, the nucleotide sequence of a single strand of DNA is shown; in reality the RNA polymerase recognizes the promoter as double-stranded DNA. On the basis of a comparison of 300 promoters, the frequencies of the four nucleotides at each position in the –35 and –10 hexamers are given. The consensus sequence, shown *below* the graph, reflects the most common nucleotide found at each position in the collection of promoters. The sequence of nucleotides between the –35 and –10 hexamers shows no significant similarities among promoters. (B) The distribution of spacing between the –35 and –10 hexamers found in *E. coli* promoters.

The information displayed in these two graphs applies to *E. coli* promoters that are recognized by RNA polymerase and the major σ factor (designated $\sigma^{70}$). As we shall see in the next chapter, bacteria also contain minor σ factors, each of which recognizes a different promoter sequence. Some particularly strong promoters recognized by RNA polymerase and $\sigma^{70}$ have an additional sequence, located upstream (to the *left,* in the figure) of the –35 hexamer, which is recognized by another subunit of RNA polymerase.

The process of transcription initiation is complex and requires that the RNA polymerase holoenzyme and the DNA undergo a series of conformational changes. We can view these changes as opening up and positioning the DNA in the active site followed by a successive tightening of the enzyme around the DNA and RNA to ensure that it does not dissociate before it has finished transcribing a gene. If an RNA polymerase does dissociate prematurely, it cannot resume synthesis but must start over again at the promoter.

How do the termination signals in the DNA stop the elongating polymerase? For most bacterial genes a termination signal consists of a string of A–T nucleotide pairs preceded by a two-fold symmetric DNA sequence, which, when transcribed into RNA, folds into a "hairpin" structure through Watson–Crick base-pairing (see Figure 6–11). As the polymerase transcribes across a terminator, the formation of the hairpin may help to "pull" the RNA transcript from the active site. The DNA–RNA hybrid in the active site, which is held together at terminators predominantly by U–A base pairs (which are less stable than G–C base pairs because they form two rather than three hydrogen bonds per base pair), is not strong enough to hold the RNA in place, and it dissociates causing the release of the polymerase from the DNA (step 7 in Figure 6–11). Thus, in some respects, transcription termination seems to involve a reversal of the structural transitions that happen during initiation. The process of termination also is an example of a common theme in this chapter: the folding of RNA into specific structures affects many steps in decoding the genome.

## Transcription Start and Stop Signals Are Heterogeneous in Nucleotide Sequence

As we have just seen, the processes of transcription initiation and termination involve a complicated series of structural transitions in protein, DNA, and RNA molecules. The signals encoded in DNA that specify these transitions are often difficult for researchers to recognize. Indeed, a comparison of many different bacterial promoters reveals a surprising degree of variation. Nevertheless, they all contain related sequences, reflecting in part aspects of the DNA that are recognized directly by the σ factor. These common features are often summarized in the form of a *consensus sequence* (**Figure 6–12**). A **consensus nucleotide sequence** is derived by comparing many sequences with the same basic function and tallying up the most common nucleotide found at each position. It

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 85 of 240

therefore serves as a summary or "average" of a large number of individual nucleotide sequences.

The DNA sequences of individual bacterial promoters differ in ways that determine their strength (the number of initiation events per unit time of the promoter). Evolutionary processes have fine-tuned each to initiate as often as necessary and have thereby created a wide spectrum of promoters. Promoters for genes that code for abundant proteins are much stronger than those associated with genes that encode rare proteins, and their nucleotide sequences are responsible for these differences.

Like bacterial promoters, transcription terminators also have a wide range of sequences, with the potential to form a simple hairpin RNA structure being the most important common feature. Since an almost unlimited number of nucleotide sequences have this potential, terminator sequences are even more heterogeneous than promoter sequences.

We have discussed bacterial promoters and terminators in some detail to illustrate an important point regarding the analysis of genome sequences. Although we know a great deal about bacterial promoters and terminators and can construct consensus sequences that summarize their most salient features, their variation in nucleotide sequence makes it difficult to definitively locate them simply by analysis of the nucleotide sequence of a genome. It is even more difficult to locate analogous sequences in eucaryotic genomes, due in part to the excess DNA carried in them. Often, we need additional information, some of it from direct experimentation, to locate and accurately interpret the short DNA signals contained in genomes.

Since DNA is double-stranded, two different RNA molecules could in principle be transcribed from any gene, using each of the two DNA strands as a template. However, a gene typically has only a single promoter, and because the promoter's nucleotide sequence is asymmetric (see Figure 6–12), the polymerase can bind in only one orientation. The polymerase synthesizes RNA in the 5′-to-3′ direction, and it can therefore only transcribe one strand per gene (**Figure 6–13**). Genome sequences reveal that the DNA strand used as the template for RNA synthesis varies from gene to gene depending on the location and orientation of the promoter (**Figure 6–14**).

Having considered transcription in bacteria, we now turn to the situation in eucaryotes, where the synthesis of RNA molecules is a much more elaborate affair.

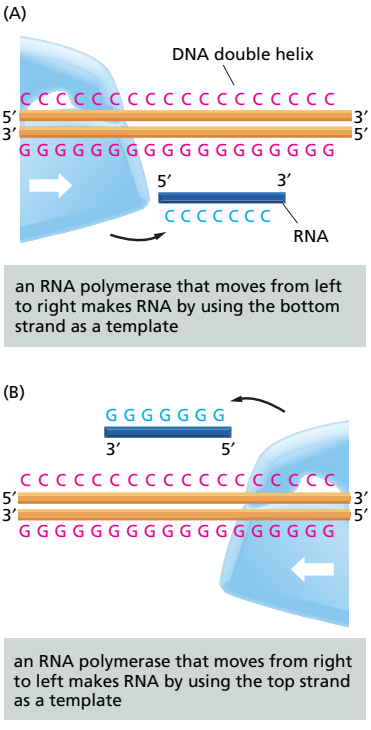



**Figure 6–13 The importance of RNA polymerase orientation.** The DNA strand serving as template must be traversed in a 3′-to-5′ direction. Thus, the direction of RNA polymerase movement determines which of the two DNA strands is to serve as a template for the synthesis of RNA, as shown in (A) and (B). Polymerase direction is, in turn, determined by the orientation of the promoter sequence, the site at which the RNA polymerase begins transcription.

## Transcription Initiation in Eucaryotes Requires Many Proteins

In contrast to bacteria, which contain a single type of RNA polymerase, eucaryotic nuclei have three: *RNA polymerase I, RNA polymerase II,* and *RNA polymerase III*. The three polymerases are structurally similar to one another (and to the bacterial enzyme) and share some common subunits, but they transcribe different types of genes (**Table 6–2**). RNA polymerases I and III transcribe the genes encoding transfer RNA, ribosomal RNA, and various small RNAs. RNA polymerase II transcribes most genes, including all those that encode proteins, and our subsequent discussion therefore focuses on this enzyme.

Although eucaryotic RNA polymerase II has many structural similarities to bacterial RNA polymerase (**Figure 6–15**), there are several important differences in the way in which the bacterial and eucaryotic enzymes function, two of which concern us immediately.

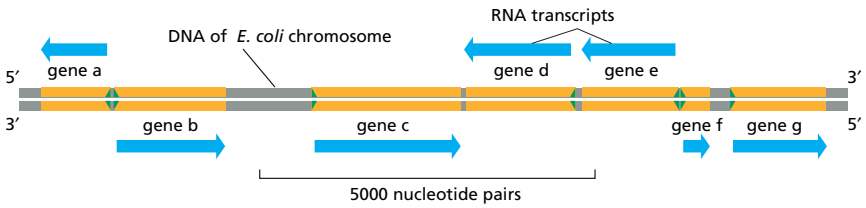



**Figure 6–14 Directions of transcription along a short portion of a bacterial chromosome.** Some genes are transcribed using one DNA strand as a template, while others are transcribed using the other DNA strand. The direction of transcription is determined by the promoter at the beginning of each gene *(green arrowheads)*. This diagram shows approximately 0.2% (9000 base pairs) of the *E. coli* chromosome. The genes transcribed from *left* to *right* use the bottom DNA strand as the template; those transcribed from *right* to *left* use the top strand as the template.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 86 of 240

**Table 6–2 The Three RNA Polymerases in Eucaryotic Cells**

| TYPE OF POLYMERASE | GENES TRANSCRIBED |
|---|---|
| RNA polymerase I | 5.8S, 18S, and 28S rRNA genes |
| RNA polymerase II | all protein-coding genes, plus snoRNA genes, miRNA genes, siRNA genes, and most snRNA genes |
| RNA polymerase III | tRNA genes, 5S rRNA genes, some snRNA genes and genes for other small RNAs |

The rRNAs are named according to their "S" values, which refer to their rate of sedimentation in an ultracentrifuge. The larger the S value, the larger the rRNA.

1. While bacterial RNA polymerase requires only a single additional protein (σ factor) for transcription initiation to occur *in vitro*, eucaryotic RNA polymerases require many additional proteins, collectively called the *general transcription factors*.
2. Eucaryotic transcription initiation must deal with the packing of DNA into nucleosomes and higher-order forms of chromatin structure, features absent from bacterial chromosomes.

## RNA Polymerase II Requires General Transcription Factors

The **general transcription factors** help to position eucaryotic RNA polymerase correctly at the promoter, aid in pulling apart the two strands of DNA to allow transcription to begin, and release RNA polymerase from the promoter into the elongation mode once transcription has begun. <CTAT> The proteins are "general" because they are needed at nearly all promoters used by RNA polymerase II; consisting of a set of interacting proteins, they are designated as *TFII* (for transcription factor for polymerase II), and are denoted arbitrarily as TFIIB, TFIID, and so on. In a broad sense, the eucaryotic general transcription factors carry out functions equivalent to those of the σ factor in bacteria; indeed, portions of TFIIF have the same three-dimensional structure as the equivalent portions of σ.

**Figure 6–16** illustrates how the general transcription factors assemble at promoters used by RNA polymerase II, and **Table 6–3** summarizes their activities. The assembly process begins when the general transcription factor TFIID binds to a short double-helical DNA sequence primarily composed of T and A nucleotides. For this reason, this sequence is known as the TATA sequence, or **TATA box**, and the subunit of TFIID that recognizes it is called TBP (for TATA-binding protein). The TATA box is typically located 25 nucleotides upstream from the transcription start site. It is not the only DNA sequence that signals the start of transcription (**Figure 6–17**), but for most polymerase II promoters it is the most important. The binding of TFIID causes a large distortion in the DNA

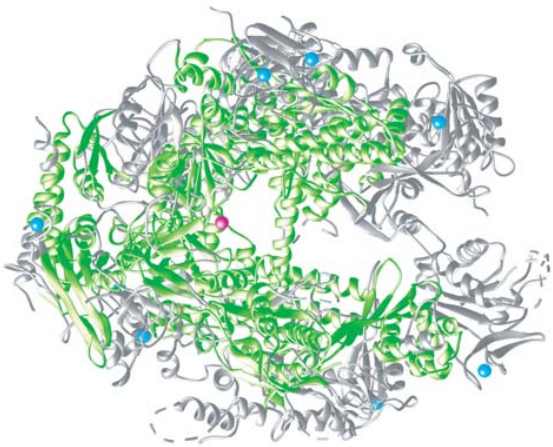


**Figure 6–15 Structural similarity between a bacterial RNA polymerase and a eucaryotic RNA polymerase II.** Regions of the two RNA polymerases that have similar structures are indicated in *green*. The eucaryotic polymerase is larger than the bacterial enzyme (12 subunits instead of 5), and some of the additional regions are shown in *gray*. The *blue* spheres represent Zn atoms that serve as structural components of the polymerases, and the *red* sphere represents the Mg atom present at the active site, where polymerization takes place. The RNA polymerases in all modern-day cells (bacteria, archaea, and eucaryotes) are closely related, indicating that the basic features of the enzyme were in place before the divergence of the three major branches of life. (Courtesy of P. Cramer and R. Kornberg.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 87 of 240

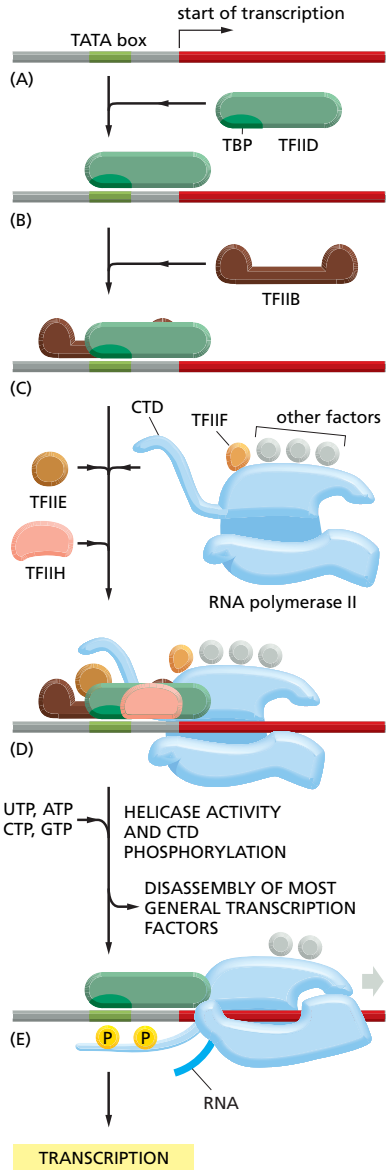



**Figure 6–16 Initiation of transcription of a eucaryotic gene by RNA polymerase II.** To begin transcription, RNA polymerase requires several general transcription factors. (A) The promoter contains a DNA sequence called the TATA box, which is located 25 nucleotides away from the site at which transcription is initiated. (B) Through its subunit TBP, TFIID recognizes and binds the TATA box, which then enables the adjacent binding of TFIIB (C). For simplicity the DNA distortion produced by the binding of TFIID (see Figure 6–18) is not shown. (D) The rest of the general transcription factors, as well as the RNA polymerase itself, assemble at the promoter. (E) TFIIH then uses ATP to pry apart the DNA double helix at the transcription start point, locally exposing the template strand. TFIIH also phosphorylates RNA polymerase II, changing its conformation so that the polymerase is released from the general factors and can begin the elongation phase of transcription. As shown, the site of phosphorylation is a long C-terminal polypeptide tail, also called the C-terminal domain (CTD), that extends from the polymerase molecule. The assembly scheme shown in the figure was deduced from experiments performed *in vitro*, and the exact order in which the general transcription factors assemble on promoters may vary from gene to gene *in vivo*. The general transcription factors have been highly conserved in evolution; some of those from human cells can be replaced in biochemical experiments by the corresponding factors from simple yeasts.

of the TATA box (**Figure 6–18**). This distortion is thought to serve as a physical landmark for the location of an active promoter in the midst of a very large genome, and it brings DNA sequences on both sides of the distortion together to allow for subsequent protein assembly steps. Other factors then assemble, along with RNA polymerase II, to form a complete *transcription initiation complex* (see Figure 6–16). The most complicated of the general transcription factors is TFIIH. Consisting of 9 subunits, it is nearly as large as RNA polymerase II itself and, as we shall see shortly, performs several enzymatic steps needed for the initiation of transcription.

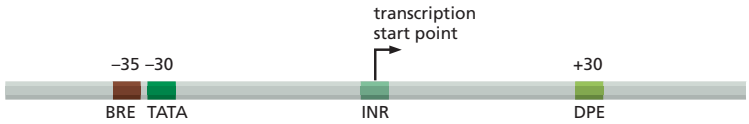
After forming a transcription initiation complex on the promoter DNA, RNA polymerase II must gain access to the template strand at the transcription start point. TFIIH, which contains a DNA helicase as one of its subunits, makes this step possible by hydrolyzing ATP and unwinding the DNA, thereby exposing the template strand. Next, RNA polymerase II, like the bacterial polymerase, remains at the promoter synthesizing short lengths of RNA until it undergoes a series of conformational changes that allow it to move away from the promoter and enter the elongation phase of transcription. A key step in this transition is the addition of phosphate groups to the "tail" of the RNA polymerase (known as the CTD or C-terminal domain). In humans, the CTD consists of 52 tandem repeats of a seven-amino-acid sequence, which extend from the RNA polymerase core structure. During transcription initiation, the serine located at the

**Table 6–3 The General Transcription Factors Needed for Transcription Initiation by Eucaryotic RNA Polymerase II**

| NAME | NUMBER OF SUBUNITS | ROLES IN TRANSITION INITIATION |
|---|---|---|
| TFIID | | |
| TBP subunit | 1 | recognizes TATA box |
| TAF subunits | ~11 | recognizes other DNA sequences near the transcription start point; regulates DNA-binding by TBP |
| TFIIB | 1 | recognizes BRE element in promoters; accurately positions RNA polymerase at the start site of transcription |
| TFIIF | 3 | stabilizes RNA polymerase interaction with TBP and TFIIB; helps attract TFIIE and TFIIH |
| TFIIE | 2 | attracts and regulates TFIIH |
| TFIIH | 9 | unwinds DNA at the transcription start point, phosphorylates Ser5 of the RNA polymerase CTD; releases RNA polymerase from the promoter |

TFIID is composed of TBP and ~11 additional subunits called TAFs (TBP-associated factors); CTD, C-terminal domain.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 88 of 240




| element | consensus sequence | general transcription factor |
|---|---|---|
| BRE | G/C G/C G/A C G C C | TFIIB |
| TATA | T A T A A/T A A/T | TBP |
| INR | C/T C/T A N T/A C/T C/T | TFIID |
| DPE | A/G G A/T C G T G | TFIID |

**Figure 6–17 Consensus sequences found in the vicinity of eucaryotic RNA polymerase II start points.** The name given to each consensus sequence *(first column)* and the general transcription factor that recognizes it *(last column)* are indicated. N indicates any nucleotide, and two nucleotides separated by a slash indicate an equal probability of either nucleotide at the indicated position. In reality, each consensus sequence is a shorthand representation of a histogram similar to that of Figure 6–12.

For most RNA polymerase II transcription start points, only two or three of the four sequences are present. For example, many polymerase II promoters have a TATA box sequence, but those that do not typically have a "strong" INR sequence. Although most of the DNA sequences that influence transcription initiation are located upstream of the transcription start point, a few, such as the DPE shown in the figure, are located in the transcribed region.

fifth position in the repeat sequence (Ser5) is phosphorylated by TFIIH, which contains a protein kinase in another of its subunits (see Figure 6–16D and E). The polymerase can then disengage from the cluster of general transcription factors. During this process, it undergoes a series of conformational changes that tighten its interaction with DNA, and it acquires new proteins that allow it to transcribe for long distances, and in some cases for many hours, without dissociating from DNA.

Once the polymerase II has begun elongating the RNA transcript, most of the general transcription factors are released from the DNA so that they are available to initiate another round of transcription with a new RNA polymerase molecule. As we see shortly, the phosphorylation of the tail of RNA polymerase II also causes components of the RNA-processing machinery to load onto the polymerase and thus be positioned to modify the newly transcribed RNA as it emerges from the polymerase.

## Polymerase II Also Requires Activator, Mediator, and Chromatin-Modifying Proteins

Studies of the behavior of RNA polymerase II and its general transcription factors on purified DNA templates *in vitro* established the model for transcription initiation just described. However, as discussed in Chapter 4, DNA in eucaryotic cells is packaged into nucleosomes, which are further arranged in higher-order

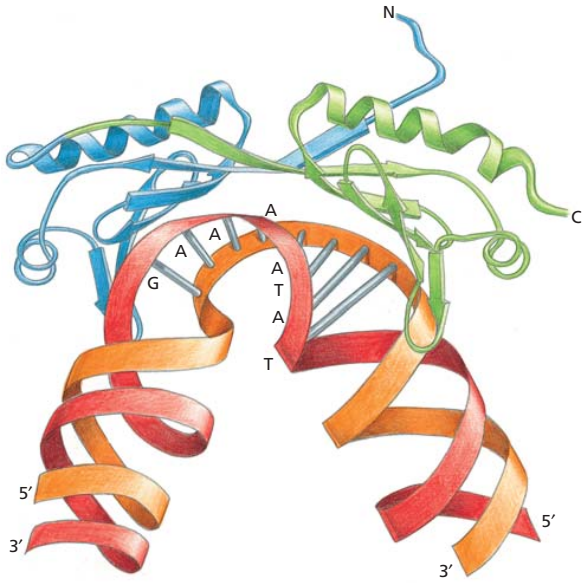



**Figure 6–18 Three-dimensional structure of TBP (TATA-binding protein) bound to DNA.** The TBP is the subunit of the general transcription factor TFIID that is responsible for recognizing and binding to the TATA box sequence in the DNA *(red)*. The unique DNA bending caused by TBP—two kinks in the double helix separated by partly unwound DNA—may serve as a landmark that helps to attract the other general transcription factors. TBP is a single polypeptide chain that is folded into two very similar domains *(blue* and *green)*. (Adapted from J.L. Kim et al., *Nature* 365:520–527, 1993. With permission from Macmillan Publishers Ltd.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 89 of 240

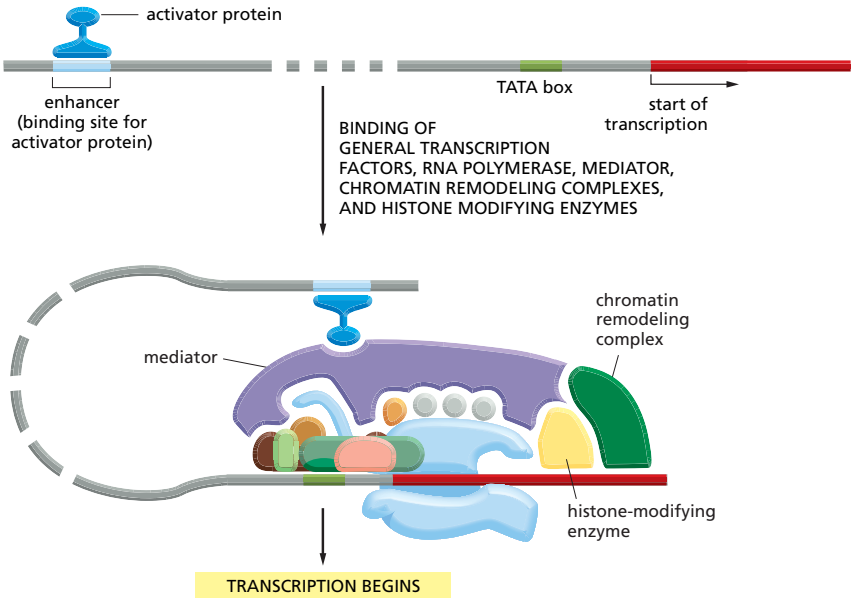



**Figure 6–19 Transcription initiation by RNA polymerase II in a eucaryotic cell.** Transcription initiation *in vivo* requires the presence of transcriptional activator proteins. As described in Chapter 7, these proteins bind to specific short sequences in DNA. Although only one is shown here, a typical eucaryotic gene has many activator proteins, which together determine its rate and pattern of transcription. Sometimes acting from a distance of several thousand nucleotide pairs (indicated by the dashed DNA molecule), these gene regulatory proteins help RNA polymerase, the general transcription factors, and the mediator all to assemble at the promoter. In addition, activators attract ATP-dependent chromatin remodeling complexes and histone acetylases.

As discussed in Chapter 4, the "default" state of chromatin is probably the 30-nm filament (see Figure 4–22), and this is likely to be a form of DNA upon which transcription is initiated. For simplicity, it is not shown in the figure.

chromatin structures. As a result, transcription initiation in a eucaryotic cell is more complex and requires even more proteins than it does on purified DNA. First, gene regulatory proteins known as *transcriptional activators* must bind to specific sequences in DNA and help to attract RNA polymerase II to the start point of transcription (**Figure 6–19**). We discuss the role of activators in Chapter 7, because they are one of the main ways in which cells regulate expression of their genes. Here we simply note that their presence on DNA is required for transcription initiation in a eucaryotic cell. Second, eucaryotic transcription initiation *in vivo* requires the presence of a protein complex known as *Mediator*, which allows the activator proteins to communicate properly with the polymerase II and with the general transcription factors. Finally, transcription initiation in a eucaryotic cell typically requires the local recruitment of chromatin-modifying enzymes, including chromatin remodeling complexes and histone-modifying enzymes. As discussed in Chapter 4, both types of enzymes can allow greater access to the DNA present in chromatin, and by doing so, they facilitate the assembly of the transcription initiation machinery onto DNA. We will revisit the role of these enzymes in transcription initiation in Chapter 7.

As illustrated in Figure 6–19, many proteins (well over 100 individual subunits) must assemble at the start point of transcription to initiate transcription in a eucaryotic cell. The order of assembly of these proteins does not seem to follow a prescribed pathway; rather, the order differs from gene to gene. Indeed, some of these different protein complexes may interact with each other away from the DNA and be brought to DNA as preformed subassemblies. To begin transcribing, RNA polymerase II must be released from this large complex of proteins, and, in addition to the steps described in Figure 6–16, this often requires the *in situ* proteolysis of the activator protein. We return to some of these issues in Chapter 7, where we discuss how eucaryotic cells can regulate the process of transcription initiation.

## Transcription Elongation Produces Superhelical Tension in DNA

Once it has initiated transcription, RNA polymerase does not proceed smoothly along a DNA molecule; rather, it moves jerkily, pausing at some sequences and rapidly transcribing through others. Elongating RNA polymerases, both bacterial and eucaryotic, are associated with a series of *elongation factors*, proteins that decrease the likelihood that RNA polymerase will dissociate before it reaches the end of a gene. These factors typically associate with RNA polymerase

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 90 of 240

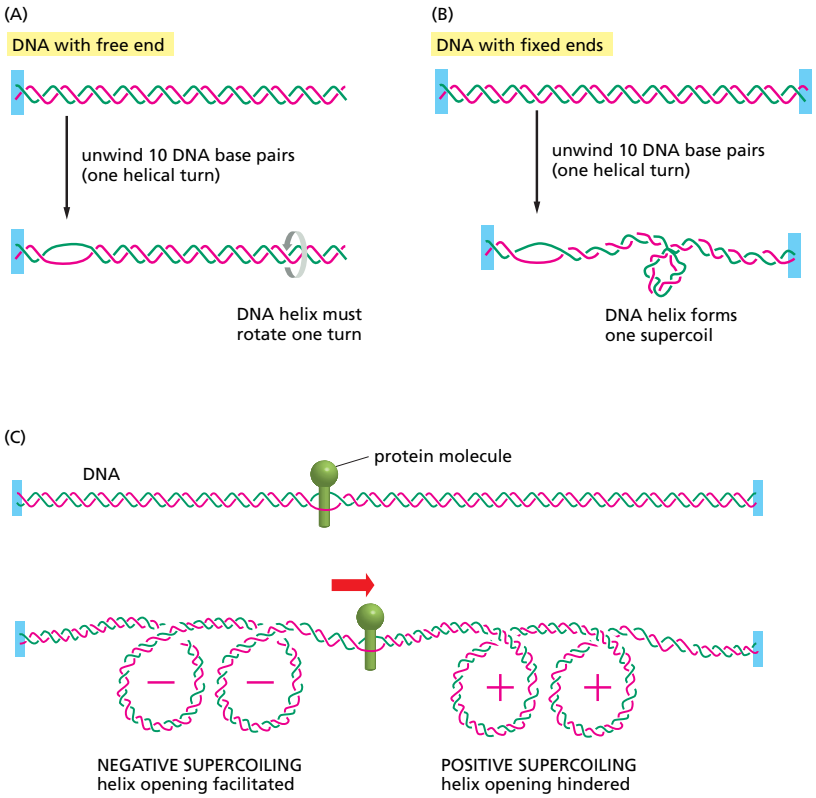



**Figure 6–20 Superhelical tension in DNA causes DNA supercoiling.** (A) For a DNA molecule with one free end (or a nick in one strand that serves as a swivel), the DNA double helix rotates by one turn for every 10 nucleotide pairs opened. (B) If rotation is prevented, superhelical tension is introduced into the DNA by helix opening. One way of accommodating this tension would be to increase the helical twist from 10 to 11 nucleotide pairs per turn in the double helix that remains; the DNA helix, however, resists such a deformation in a springlike fashion, preferring to relieve the superhelical tension by bending into supercoiled loops. As a result, one DNA supercoil forms in the DNA double helix for every 10 nucleotide pairs opened. The supercoil formed in this case is a positive supercoil. (C) Supercoiling of DNA is induced by a protein tracking through the DNA double helix. The two ends of the DNA shown here are unable to rotate freely relative to each other, and the protein molecule is assumed also to be prevented from rotating freely as it moves. Under these conditions, the movement of the protein causes an excess of helical turns to accumulate in the DNA helix ahead of the protein and a deficit of helical turns to arise in the DNA behind the protein, as shown.

shortly after initiation and help polymerases to move through the wide variety of different DNA sequences that are found in genes. Eucaryotic RNA polymerases must also contend with chromatin structure as they move along a DNA template, and they are typically aided by ATP-dependent chromatin remodeling complexes (see pp. 215–216). These complexes may move with the polymerase or may simply seek out and rescue the occasional stalled polymerase. In addition, some elongation factors associated with eucaryotic RNA polymerase facilitate transcription through nucleosomes without requiring additional energy. It is not yet understood in detail how this is accomplished, but these proteins can transiently dislodge H2A–H2B dimers from the nucleosome core, replacing them as the polymerase moves through the nucleosome.

There is yet another barrier to elongating polymerases, both bacterial and eucaryotic. To discuss this issue, we need first to consider a subtle property inherent in the DNA double helix called **DNA supercoiling**. DNA supercoiling represents a conformation that DNA adopts in response to superhelical tension; conversely, creating various loops or coils in the helix can create such tension. **Figure 6–20** illustrates the topological constraints that cause DNA supercoiling. There are approximately 10 nucleotide pairs for every helical turn in a DNA double helix. Imagine a helix whose two ends are fixed with respect to each other (as they are in a DNA circle, such as a bacterial chromosome, or in a tightly clamped loop, as is thought to exist in eucaryotic chromosomes). In this case, one large DNA supercoil will form to compensate for each 10 nucleotide pairs that are opened (unwound). The formation of this supercoil is energetically favorable because it restores a normal helical twist to the base-paired regions that remain, which would otherwise need to be overwound because of the fixed ends.

RNA polymerase also creates superhelical tension as it moves along a stretch of DNA that is anchored at its ends (see Figure 6–20C). As long as the polymerase is not free to rotate rapidly (and such rotation is unlikely given the size of RNA polymerases and their attached transcripts), a moving polymerase generates

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 91 of 240

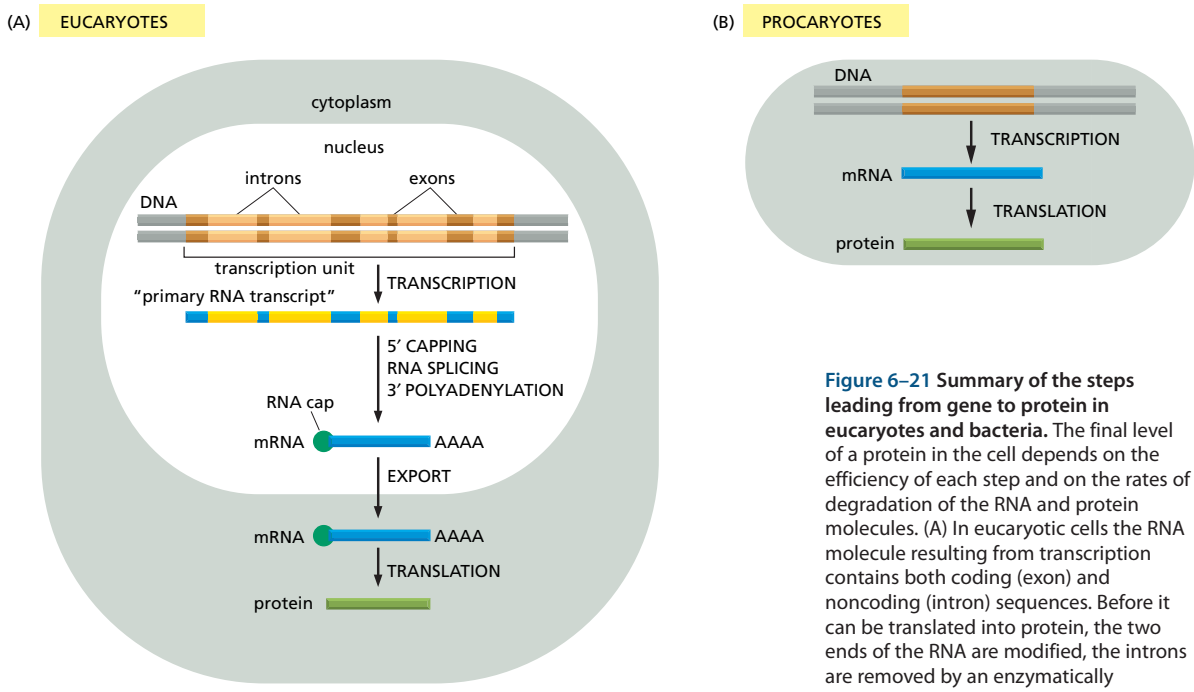



**Figure 6–21 Summary of the steps leading from gene to protein in eucaryotes and bacteria.** The final level of a protein in the cell depends on the efficiency of each step and on the rates of degradation of the RNA and protein molecules. (A) In eucaryotic cells the RNA molecule resulting from transcription contains both coding (exon) and noncoding (intron) sequences. Before it can be translated into protein, the two ends of the RNA are modified, the introns are removed by an enzymatically catalyzed RNA splicing reaction, and the resulting mRNA is transported from the nucleus to the cytoplasm. Although the steps in this figure are depicted as occurring one at a time, in a sequence, in reality they can occur concurrently. For example, the RNA cap is added and splicing typically begins before transcription has been completed. Because of the coupling between transcription and RNA processing, primary transcripts—the RNAs that would, in theory, be produced if no processing had occurred—are found only rarely. (B) In procaryotes the production of mRNA is much simpler. The 5′ end of an mRNA molecule is produced by the initiation of transcription, and the 3′ end is produced by the termination of transcription. Since procaryotic cells lack a nucleus, transcription and translation take place in a common compartment. In fact, the translation of a bacterial mRNA often begins before its synthesis has been completed.

positive superhelical tension in the DNA in front of it and negative helical tension behind it. For eucaryotes, this situation is thought to provide a bonus: the positive superhelical tension ahead of the polymerase makes the DNA helix more difficult to open, but this tension should facilitate the unwrapping of DNA in nucleosomes, as the release of DNA from the histone core helps to relax positive superhelical tension.

Any protein that propels itself alone along a DNA strand of a double helix tends to generate superhelical tension. In eucaryotes, DNA topoisomerase enzymes rapidly remove this superhelical tension (see p. 278). But in bacteria a specialized topoisomerase called *DNA gyrase* uses the energy of ATP hydrolysis to pump supercoils continuously into the DNA, thereby maintaining the DNA under constant tension. These are *negative supercoils*, having the opposite handedness from the *positive supercoils* that form when a region of DNA helix opens (see Figure 6–20B). Whenever a region of helix opens, it removes these negative supercoils from bacterial DNA, reducing the superhelical tension. DNA gyrase therefore makes the opening of the DNA helix in bacteria energetically favorable compared with helix opening in DNA that is not supercoiled. For this reason, it usually facilitates those genetic processes in bacteria, including the initiation of transcription by bacterial RNA polymerase, that require helix opening (see Figure 6–11).

## Transcription Elongation in Eucaryotes Is Tightly Coupled to RNA Processing

We have seen that bacterial mRNAs are synthesized solely by the RNA polymerase starting and stopping at specific spots on the genome. The situation in eucaryotes is substantially different. In particular, transcription is only the first of several steps needed to produce an mRNA. Other critical steps are the covalent modification of the ends of the RNA and the removal of *intron sequences* that are discarded from the middle of the RNA transcript by the process of *RNA splicing* (**Figure 6–21**).

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 92 of 240

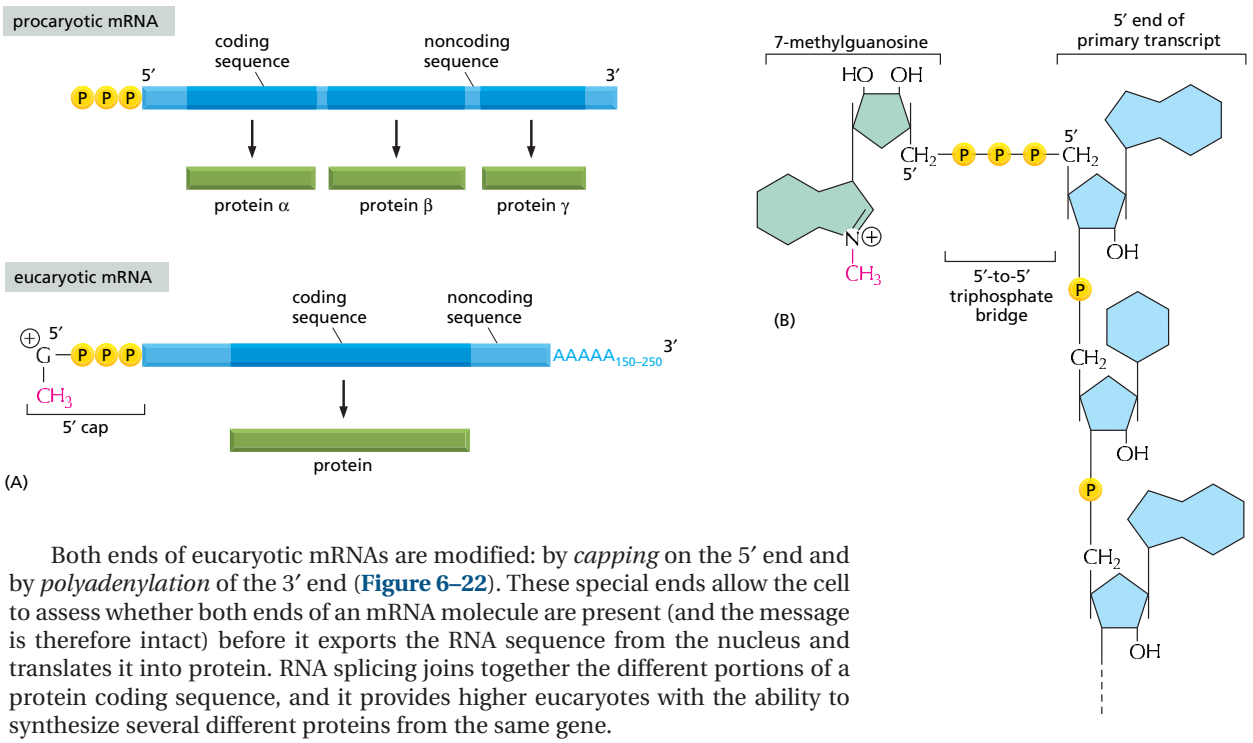



**Figure 6–22 A comparison of the structures of procaryotic and eucaryotic mRNA molecules.** (A) The 5′ and 3′ ends of a bacterial mRNA are the unmodified ends of the chain synthesized by the RNA polymerase, which initiates and terminates transcription at those points, respectively. The corresponding ends of a eucaryotic mRNA are formed by adding a 5′ cap and by cleavage of the pre-mRNA transcript and the addition of a poly-A tail, respectively. The figure also illustrates another difference between the procaryotic and eucaryotic mRNAs: bacterial mRNAs can contain the instructions for several different proteins, whereas eucaryotic mRNAs nearly always contain the information for only a single protein. (B) The structure of the cap at the 5′ end of eucaryotic mRNA molecules. Note the unusual 5′-to-5′ linkage of the 7-methyl G to the remainder of the RNA. Many eucaryotic mRNAs carry an additional modification: the 2′-hydroxyl group on the second ribose sugar in the mRNA is methylated (not shown).

Both ends of eucaryotic mRNAs are modified: by *capping* on the 5′ end and by *polyadenylation* of the 3′ end (**Figure 6–22**). These special ends allow the cell to assess whether both ends of an mRNA molecule are present (and the message is therefore intact) before it exports the RNA sequence from the nucleus and translates it into protein. RNA splicing joins together the different portions of a protein coding sequence, and it provides higher eucaryotes with the ability to synthesize several different proteins from the same gene.

An ingenious mechanism couples all of the above RNA processing steps to transcription elongation. As discussed previously, a key step in transcription initiation by RNA polymerase II is the phosphorylation of the RNA polymerase II tail, called the CTD (C-terminal domain). This phosphorylation proceeds gradually as the RNA polymerase initiates transcription and moves along the DNA. It not only helps dissociate the RNA polymerase II from other proteins present at the start point of transcription, but also allows a new set of proteins to associate with the RNA polymerase tail that function in transcription elongation and RNA processing. As discussed next, some of these processing proteins seem to "hop" from the polymerase tail onto the nascent RNA molecule to begin processing it as it emerges from the RNA polymerase. Thus, we can view RNA polymerase II in its elongation mode as an RNA factory that both transcribes DNA into RNA and processes the RNA it produces (**Figure 6–23**). Fully extended, the CTD is nearly 10 times longer than the remainder of RNA polymerase and, in effect, it serves as a tether, holding a variety of proteins close by until they are needed. This strategy, which speeds up the rate of subsequent reactions, is one commonly observed in the cell (see Figures 4–69 and 16–38).

## RNA Capping Is the First Modification of Eucaryotic Pre-mRNAs

As soon as RNA polymerase II has produced about 25 nucleotides of RNA, the 5′ end of the new RNA molecule is modified by addition of a cap that consists of a modified guanine nucleotide (see Figure 6–22B). Three enzymes, acting in succession, perform the capping reaction: one (a phosphatase) removes a phosphate from the 5′ end of the nascent RNA, another (a guanyl transferase) adds a GMP in a reverse linkage (5′ to 5′ instead of 5′ to 3′), and a third (a methyl transferase) adds a methyl group to the guanosine (**Figure 6–24**). Because all three enzymes bind to the RNA polymerase tail phosphorylated at serine-5 position, the modification added by TFIIH during transcription initiation, they are poised to modify the 5′ end of the nascent transcript as soon as it emerges from the polymerase.

The 5′-methyl cap signifies the 5′ end of eucaryotic mRNAs, and this landmark helps the cell to distinguish mRNAs from the other types of RNA molecules present in the cell. For example, RNA polymerases I and III produce uncapped

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 93 of 240

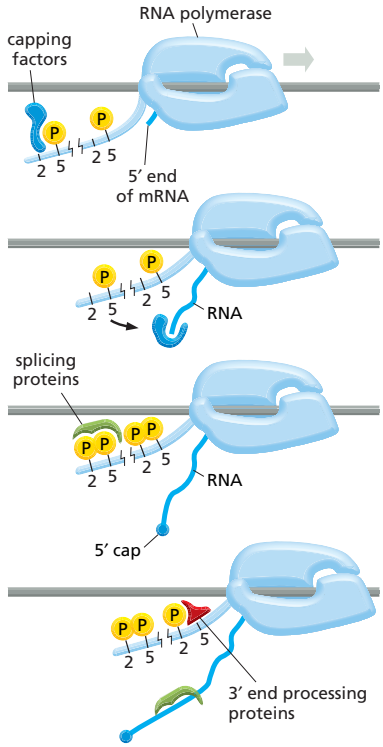



**Figure 6–23 Eucaryotic RNA polymerase II as an "RNA factory."** As the polymerase transcribes DNA into RNA, it carries pre-mRNA-processing proteins on its tail that are transferred to the nascent RNA at the appropriate time. The tail, known as the CTD, contains 52 tandem repeats of a seven amino acid sequence, and there are two serines in each repeat. The capping proteins first bind to the RNA polymerase tail when it is phosphorylated on Ser5 of the heptad repeat late in the process of transcription initiation (see Figure 6–16). This strategy ensures that the RNA molecule is efficiently capped as soon as its 5′ end emerges from the RNA polymerase. As the polymerase continues transcribing, its tail is extensively phosphorylated on the Ser2 positions by a kinase associated with the elongating polymerase and is eventually dephosphorylated at Ser5 positions. These further modifications attract splicing and 3′-end processing proteins to the moving polymerase, positioning them to act on the newly synthesized RNA as it emerges from the RNA polymerase. There are many RNA-processing enzymes, and not all travel with the polymerase. For RNA splicing, for example, the tail carries only a few critical components; once transferred to an RNA molecule, they serve as a nucleation site for the remaining components.

When RNA polymerase II finishes transcribing a gene, it is released from DNA, soluble phosphatases remove the phosphates on its tail, and it can reinitiate transcription. Only the dephosphorylated form of RNA polymerase II is competent to begin RNA synthesis at a promoter.

RNAs during transcription, in part because these polymerases lack a CTD. In the nucleus, the cap binds a protein complex called CBC (cap-binding complex), which, as we discuss in subsequent sections, helps the RNA to be properly processed and exported. The 5′-methyl cap also has an important role in the translation of mRNAs in the cytosol, as we discuss later in the chapter.

## RNA Splicing Removes Intron Sequences from Newly Transcribed Pre-mRNAs

As discussed in Chapter 4, the protein coding sequences of eucaryotic genes are typically interrupted by noncoding intervening sequences (introns). Discovered in 1977, this feature of eucaryotic genes came as a surprise to scientists, who had been, until that time, familiar only with bacterial genes, which typically consist of a continuous stretch of coding DNA that is directly transcribed into mRNA. In marked contrast, eucaryotic genes were found to be broken up into small pieces of coding sequence (*expressed sequences* or **exons**) interspersed with much longer *intervening sequences* or **introns**; thus, the coding portion of a eucaryotic gene is often only a small fraction of the length of the gene (**Figure 6–25**).

Both intron and exon sequences are transcribed into RNA. The intron sequences are removed from the newly synthesized RNA through the process of **RNA splicing**. The vast majority of RNA splicing that takes place in cells functions in the production of mRNA, and our discussion of splicing focuses on this so-called precursor-mRNA (or pre-mRNA) splicing. Only after 5′ and 3′ end processing and splicing have taken place is such RNA termed mRNA.

Each splicing event removes one intron, proceeding through two sequential phosphoryl-transfer reactions known as transesterifications; these join two exons while removing the intron as a "lariat" (**Figure 6–26**). Since the number of high-energy phosphate bonds remains the same, these reactions could in

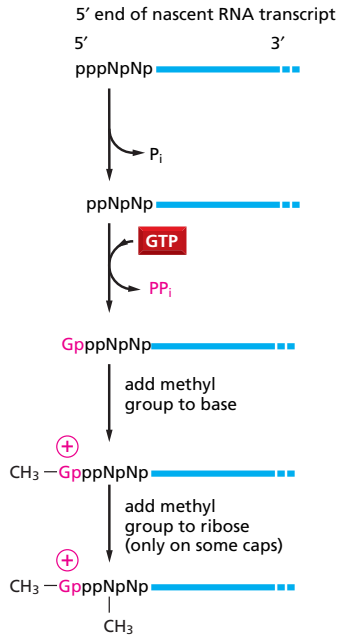



**Figure 6–24 The reactions that cap the 5′ end of each RNA molecule synthesized by RNA polymerase II.** The final cap contains a novel 5′-to-5′ linkage between the positively charged 7-methyl G residue and the 5′ end of the RNA transcript (see Figure 6–22B). The letter N represents any one of the four ribonucleotides, although the nucleotide that starts an RNA chain is usually a purine (an A or a G). (After A.J. Shatkin, *BioEssays* 7:275–277, 1987. With permission from ICSU Press.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 94 of 240

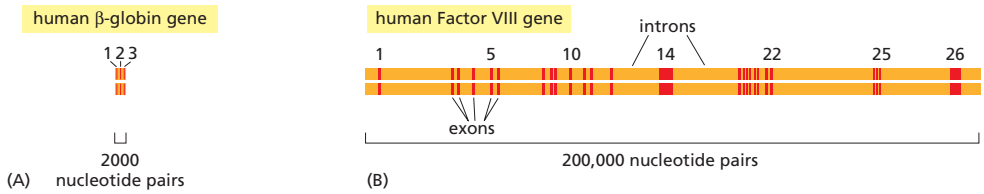



**Figure 6–25 Structure of two human genes showing the arrangement of exons and introns.** (A) The relatively small β-globin gene, which encodes one of the subunits of the oxygen-carrying protein hemoglobin, contains 3 exons (see also Figure 4–7). (B) The much larger Factor VIII gene contains 26 exons; it codes for a protein (Factor VIII) that functions in the blood-clotting pathway. The most prevalent form of hemophilia results from mutations in this gene.

principle take place without nucleoside triphosphate hydrolysis. However, the machinery that catalyzes pre-mRNA splicing is complex, consisting of 5 additional RNA molecules and as many as 200 proteins, and it hydrolyzes many ATP molecules per splicing event. This additional complexity ensures that splicing is accurate, while at the same time being flexible enough to deal with the enormous variety of introns found in a typical eucaryotic cell.

It may seem wasteful to remove large numbers of introns by RNA splicing. In attempting to explain why it occurs, scientists have pointed out that the exon–intron arrangement would seem to facilitate the emergence of new and useful proteins over evolutionary time scales. Thus, the presence of numerous introns in DNA allows genetic recombination to readily combine the exons of different genes (see p. 140), enabling genes for new proteins to evolve more easily by the combination of parts of preexisting genes. The observation, described in Chapter 3, that many proteins in present-day cells resemble patchworks composed from a common set of protein *domains*, supports this idea.

RNA splicing also has a present-day advantage. The transcripts of many eucaryotic genes (estimated at 75% of genes in humans) are spliced in more than one way, thereby allowing the same gene to produce a corresponding set of different proteins (**Figure 6–27**). Rather than being the wasteful process it may have seemed at first sight, RNA splicing enables eucaryotes to increase the already enormous coding potential of their genomes. We shall return to this idea again in this chapter and the next, but we first need to describe the cellular machinery that performs this remarkable task.

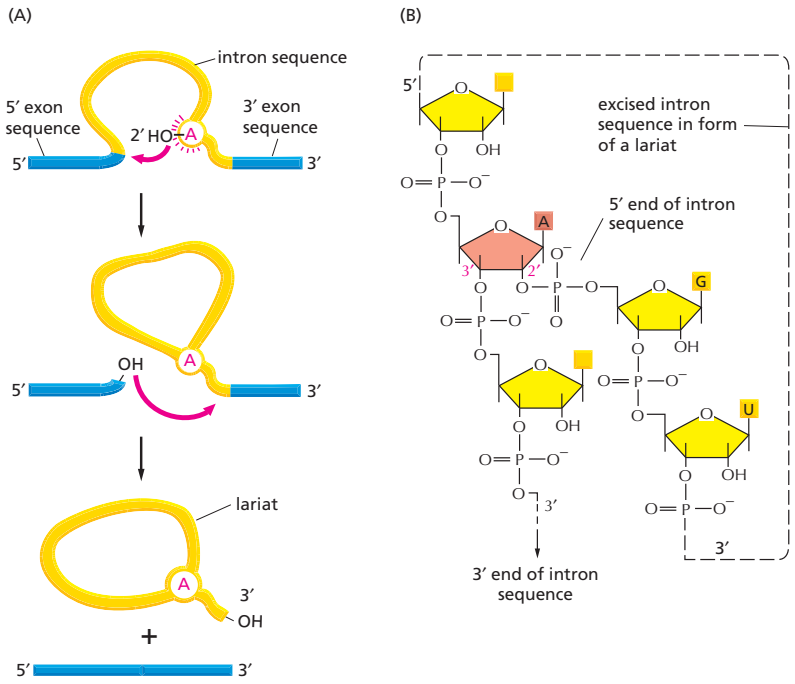



**Figure 6–26 The pre-mRNA splicing reaction.** (A) In the first step, a specific adenine nucleotide in the intron sequence (indicated in *red)* attacks the 5′ splice site and cuts the sugar-phosphate backbone of the RNA at this point. The cut 5′ end of the intron becomes covalently linked to the adenine nucleotide, as shown in detail in (B), thereby creating a loop in the RNA molecule. The released free 3′-OH end of the exon sequence then reacts with the start of the next exon sequence, joining the two exons together and releasing the intron sequence in the shape of a *lariat.* The two exon sequences thereby become joined into a continuous coding sequence; the released intron sequence is eventually degraded.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 95 of 240

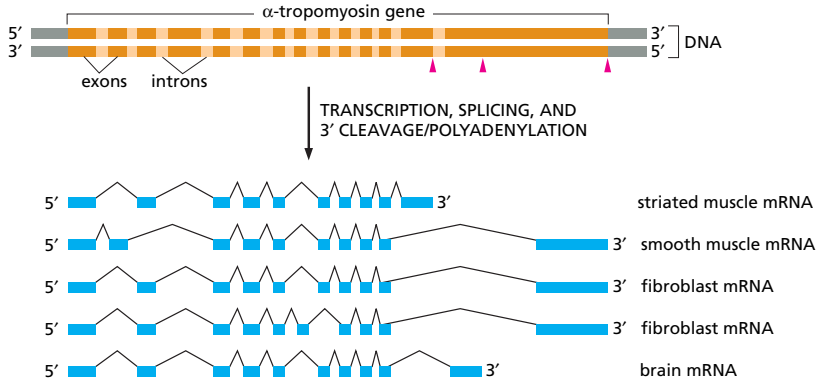



**Figure 6–27 Alternative splicing of the α-tropomyosin gene from rat.** α-Tropomyosin is a coiled-coil protein (see Figure 3–9) that regulates contraction in muscle cells. The primary transcript can be spliced in different ways, as indicated in the figure, to produce distinct mRNAs, which then give rise to variant proteins. Some of the splicing patterns are specific for certain types of cells. For example, the α-tropomyosin made in striated muscle is different from that made from the same gene in smooth muscle. The arrowheads in the top part of the figure mark the sites where cleavage and poly-A addition form the 3′ ends of the mature mRNAs.

## Nucleotide Sequences Signal Where Splicing Occurs

The mechanism of pre-mRNA splicing shown in Figure 6–26 implies that the splicing machinery must recognize three portions of the precursor RNA molecule: the 5′ splice site, the 3′ splice site, and the branch point in the intron sequence that forms the base of the excised lariat. Not surprisingly, each site has a consensus nucleotide sequence that is similar from intron to intron and provides the cell with cues for where splicing is to take place (**Figure 6–28**). However, these consensus sequences are relatively short and can accommodate a high degree of sequence variability; as we shall see shortly, the cell incorporates additional types of information to ultimately choose exactly where, on each RNA molecule, splicing is to take place.

The high variability of the splicing consensus sequences presents a special challenge for scientists attempting to decipher genome sequences. Introns range in size from about 10 nucleotides to over 100,000 nucleotides, and choosing the precise borders of each intron is a difficult task even with the aid of powerful computers. The possibility of alternative splicing compounds the problem of predicting protein sequences solely from a genome sequence. This difficulty is one of the main barriers to identifying all of the genes in a complete genome sequence, and it is one of the primary reasons why we know only the approximate number of genes in the human genome.

## RNA Splicing Is Performed by the Spliceosome

Unlike the other steps of mRNA production we have discussed, key steps in RNA splicing are performed by RNA molecules rather than proteins. Specialized RNA molecules recognize the nucleotide sequences that specify where splicing is to occur and also participate in the chemistry of splicing. These RNA molecules are relatively short (less than 200 nucleotides each), and there are five of them (U1, U2, U4, U5, and U6) involved in the major form of pre-mRNA splicing. Known as

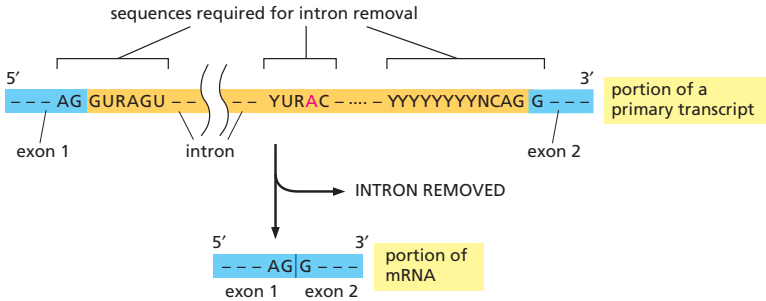


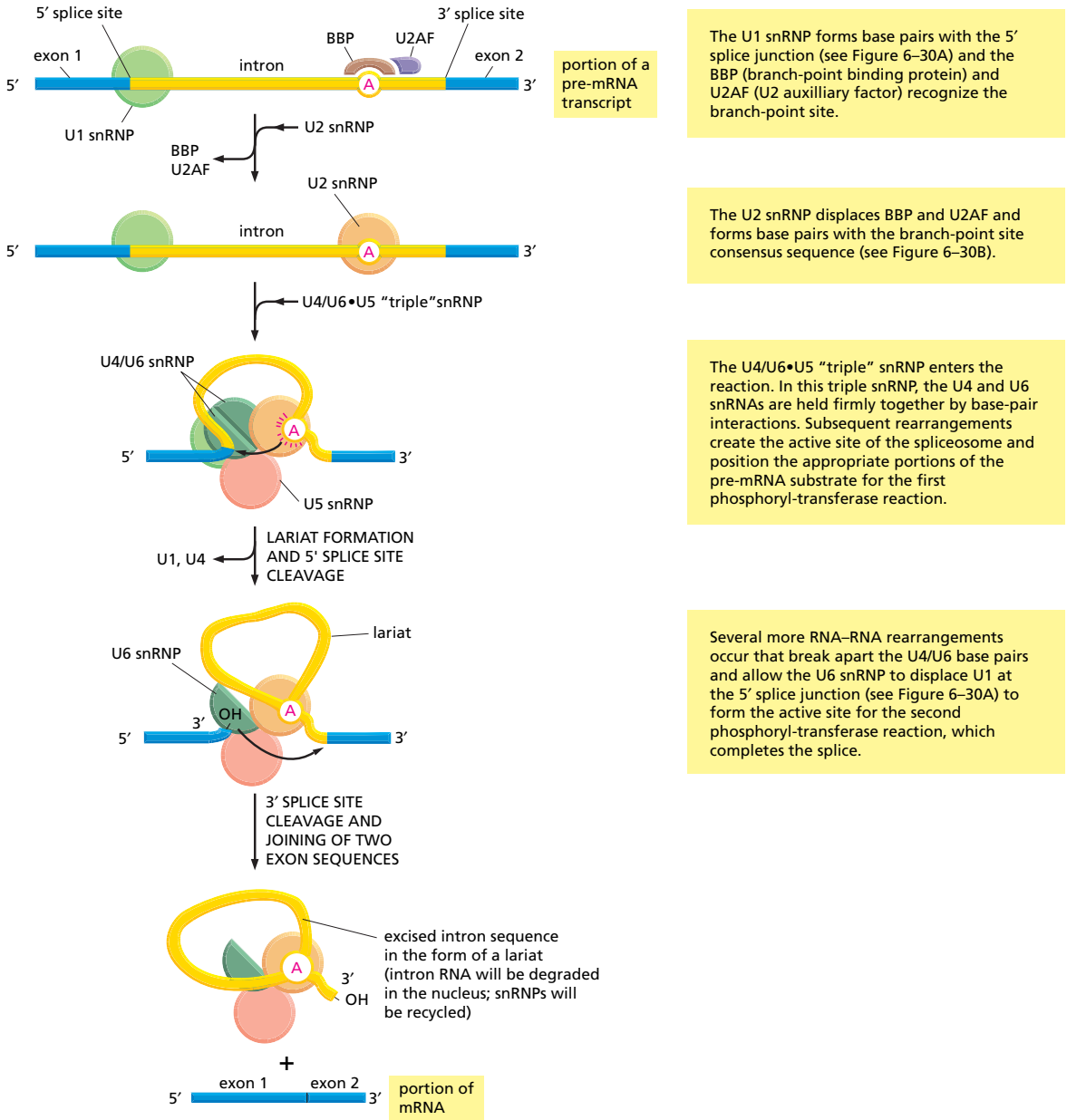

**Figure 6–28 The consensus nucleotide sequences in an RNA molecule that signal the beginning and the end of most introns in humans.** Only the three blocks of nucleotide sequences shown are required to remove an intron sequence; the rest of the intron can be occupied by any nucleotides. Here A, G, U, and C are the standard RNA nucleotides; R stands for purines (A or G); and Y stands for pyrimidines (C or U). The A highlighted in *red* forms the branch point of the lariat produced by splicing. Only the GU at the start of the intron and the AG at its end are invariant nucleotides in the splicing consensus sequences. Several different nucleotides can occupy the remaining positions (even the branch point A), although the indicated nucleotides are preferred. The distances along the RNA between the three splicing consensus sequences are highly variable; however, the distance between the branch point and 3′ splice junction is typically much shorter than that between the 5′ splice junction and the branch point.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 96 of 240

**snRNAs** (**small nuclear RNAs**), each is complexed with at least seven protein subunits to form a snRNP (small nuclear ribonucleoprotein). These snRNPs form the core of the **spliceosome**, the large assembly of RNA and protein molecules that performs pre-mRNA splicing in the cell.

The spliceosome is a complex and dynamic machine. When studied *in vitro,* a few components of the spliceosome assemble on pre-mRNA and, as the splicing reaction proceeds, new components enter as those that have already performed their tasks are jettisoned (**Figure 6–29**). However, many scientists believe that, inside the cell, the spliceosome is a preexisting, loose assembly of all the components—capturing, splicing and releasing RNA as a coordinated unit, and undergoing extensive rearrangements each time a splice is made. During the splicing reaction, recognition of the 5′ splice junction, the branch-point site, and the 3′ splice junction is performed largely through base-pairing between the snRNAs and the consensus RNA sequences in the pre-mRNA substrate (**Figure**

**Figure 6–29 The pre-mRNA splicing mechanism.** RNA splicing is catalyzed by an assembly of snRNPs (shown as *colored circles)* plus other proteins (most of which are not shown), which together constitute the spliceosome. The spliceosome recognizes the splicing signals on a pre-mRNA molecule, brings the two ends of the intron together, and provides the enzymatic activity for the two reaction steps (see Figure 6–26).




LCY Biotechnology Holding, Inc.
Ex. 1008
Page 97 of 240

6–30). In the course of splicing, the spliceosome undergoes several shifts in which one set of base-pair interactions is broken and another is formed in its place. For example, U1 is replaced by U6 at the 5′ splice junction (see Figure 6–30A). This type of RNA–RNA rearrangement (in which the formation of one RNA–RNA interaction requires the disruption of another) occurs several times during the splicing reaction. It permits the checking and rechecking of RNA sequences before the chemical reaction is allowed to proceed, thereby increasing the accuracy of splicing.

## The Spliceosome Uses ATP Hydrolysis to Produce a Complex Series of RNA–RNA Rearrangements

Although ATP hydrolysis is not required for the chemistry of RNA splicing *per se*, it is required for the assembly and rearrangements of the spliceosome. Some of the additional proteins that make up the spliceosome use the energy of ATP hydrolysis to break existing RNA–RNA interactions to allow the formation of new ones. In fact, all the steps shown previously in Figure 6–29—except the association of BBP with the branch-point site and U1 snRNP with the 5′ splice site—require ATP hydrolysis and additional proteins. Each successful splice requires as many as 200 proteins, if we include those that form the snRNPs.

The ATP-requiring RNA–RNA rearrangements that take place in the spliceosome occur within the snRNPs themselves and between the snRNPs and the pre-mRNA substrate. One of the most important functions of these rearrangements is the creation of the active catalytic site of the spliceosome. The strategy of creating an active site only after the assembly and rearrangement of splicing components on a pre-mRNA substrate is a particularly effective way to prevent wayward splicing.

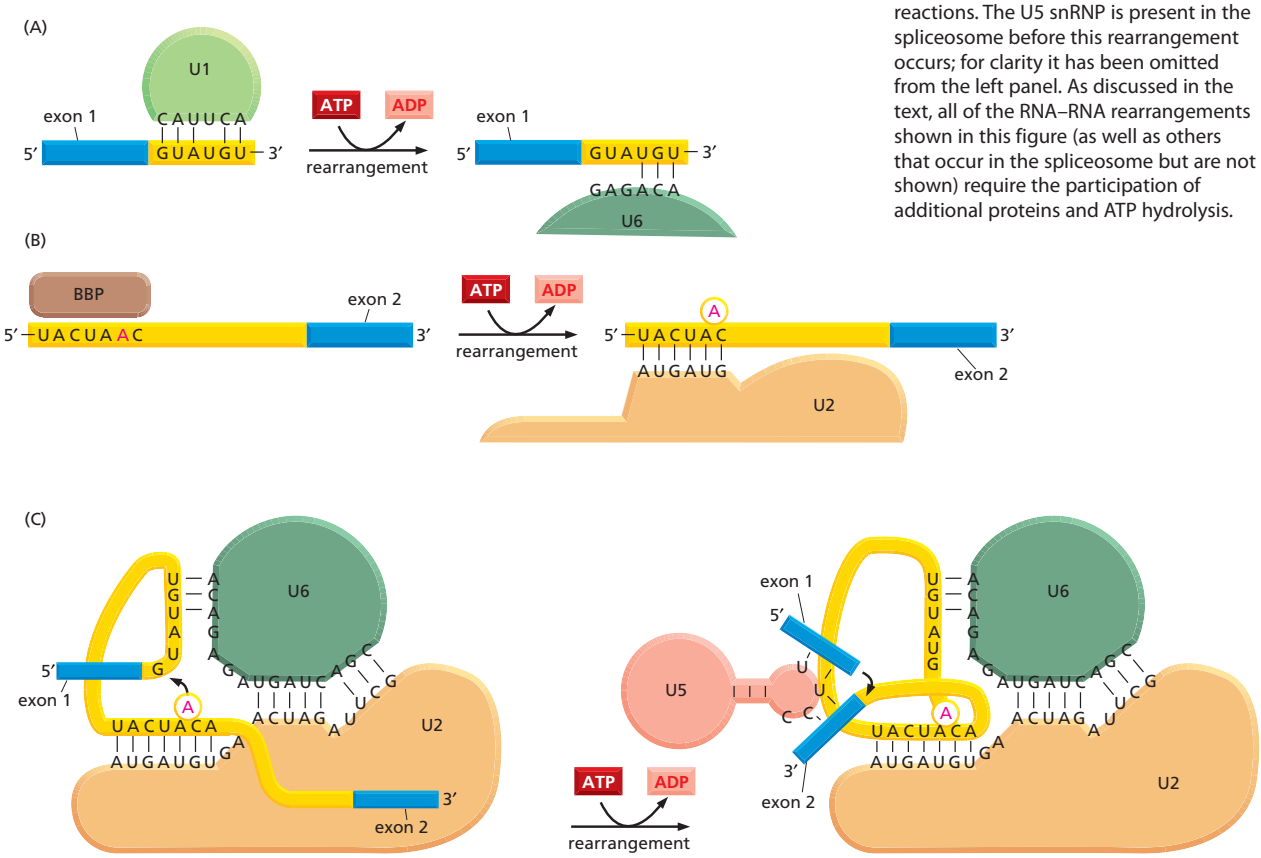



**Figure 6–30 Several of the rearrangements that take place in the spliceosome during pre-mRNA splicing.** Shown here are the details for the yeast *Saccharomyces cerevisiae*, in which the nucleotide sequences involved are slightly different from those in human cells. (A) The exchange of U1 snRNP for U6 snRNP occurs before the first phosphoryl-transfer reaction (see Figure 6–29). This exchange requires the 5′ splice site to be read by two different snRNPs, thereby increasing the accuracy of 5′ splice site selection by the spliceosome. (B) The branch-point site is first recognized by BBP and subsequently by U2 snRNP; as in (A), this "check and recheck" strategy provides increased accuracy of site selection. The binding of U2 to the branch point forces the appropriate adenine (in *red*) to be unpaired and thereby activates it for the attack on the 5′ splice site (see Figure 6–29). This, in combination with recognition by BBP, is the way in which the spliceosome accurately chooses the adenine that is ultimately to form the branch point. (C) After the first phosphoryl-transfer reaction *(left)* has occurred, a series of rearrangements brings the two exons into close proximity for the second phosphoryl-transfer reaction *(right)*. The snRNAs both position the reactants and provide (either all or in part) the catalytic sites for the two reactions. The U5 snRNP is present in the spliceosome before this rearrangement occurs; for clarity it has been omitted from the left panel. As discussed in the text, all of the RNA–RNA rearrangements shown in this figure (as well as others that occur in the spliceosome but are not shown) require the participation of additional proteins and ATP hydrolysis.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 98 of 240

Perhaps the most surprising feature of the spliceosome is the nature of the catalytic site itself: it is largely (if not exclusively) formed by RNA molecules instead of proteins. In the last section of this chapter we discuss in general terms the structural and chemical properties of RNA that allow it to perform catalysis; here we need only consider that the U2 and U6 snRNAs in the spliceosome form a precise three-dimensional RNA structure that juxtaposes the 5′ splice site of the pre-mRNA with the branch-point site and probably performs the first transesterification reaction (see Figure 6–30C). In a similar way, the 5′ and 3′ splice junctions are brought together (an event requiring the U5 snRNA) to facilitate the second transesterification.

Once the splicing chemistry is completed, the snRNPs remain bound to the lariat. The disassembly of these snRNPs from the lariat (and from each other) requires another series of RNA–RNA rearrangements that require ATP hydrolysis, thereby returning the snRNAs to their original configuration so that they can be used again in a new reaction. At the completion of a splice, the spliceosome directs a set of proteins to bind to the mRNA near the position formerly occupied by the intron. Called the *exon junction complex (EJC),* these proteins mark the site of a successful splicing event and, as we shall see later in this chapter, influence the subsequent fate of the mRNA.

## Other Properties of Pre-mRNA and Its Synthesis Help to Explain the Choice of Proper Splice Sites

As we have seen, intron sequences vary enormously in size, with some being in excess of 100,000 nucleotides. If splice-site selection were determined solely by the snRNPs acting on a preformed, protein-free RNA molecule, we would expect splicing mistakes—such as exon skipping and the use of "cryptic" splice sites—to be very common (**Figure 6–31**). The fidelity mechanisms built into the spliceosome, however, are supplemented by two additional strategies that increase the accuracy of splicing. The first is simply a consequence of the early stages of splicing occurring while the pre-mRNA molecules are being synthesized by RNA polymerase II. As transcription proceeds, the phosphorylated tail of RNA polymerase carries several components of the spliceosome (see Figure 6–23), and these components are transferred directly from the polymerase to the RNA as RNA is synthesized. This strategy helps the cell keep track of introns and exons: for example, the snRNPs that assemble at a 5′ splice site are initially presented with only a single 3′ splice site since the sites further downstream have not yet been synthesized. The coordination of transcription with splicing is especially important in preventing inappropriate exon skipping.

A strategy called "exon definition" is another way cells choose the appropriate splice sites. Exon size tends to be much more uniform than intron size, averaging about 150 nucleotide pairs across a wide variety of eucaryotic organisms (**Figure 6–32**). According to the exon definition idea, the splicing machinery initially seeks out the relatively homogenously sized exon sequences. As RNA synthesis proceeds, a group of additional components (most notably SR proteins, so-named because they contain a domain rich in serines and arginines) assemble on exon sequences and help to mark off each 3′ and 5′ splice site starting at the 5′ end of the RNA (**Figure 6–33**). These proteins, in turn, recruit U1 snRNA, which

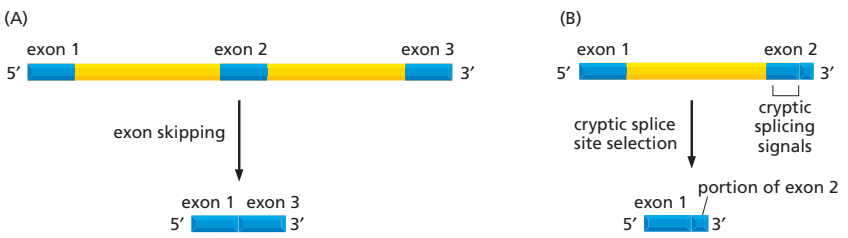



**Figure 6–31 Two types of splicing errors.** (A) Exon skipping. (B) Cryptic splice-site selection. Cryptic splicing signals are nucleotide sequences of RNA that closely resemble true splicing signals.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 99 of 240

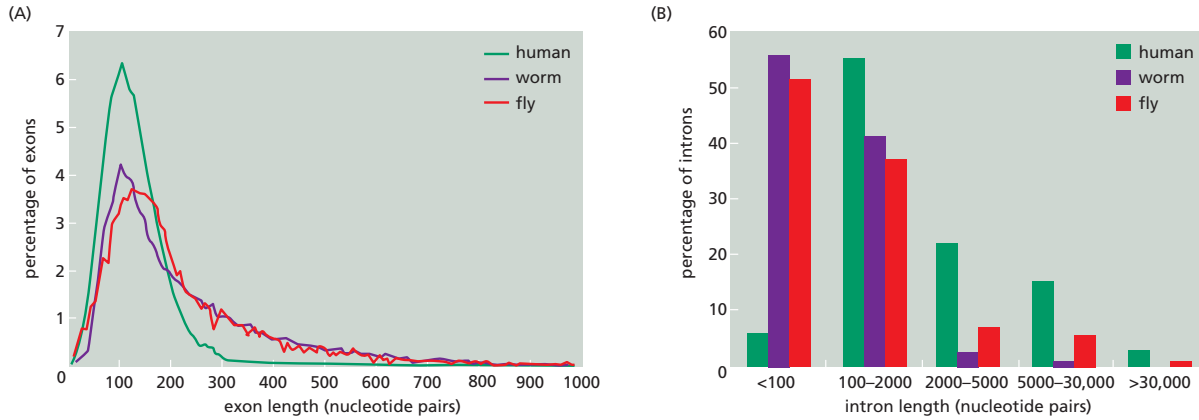



**Figure 6–32 Variation in intron and exon lengths in the human, worm, and fly genomes.** (A) Size distribution of exons. (B) Size distribution of introns. Note that exon length is much more uniform than intron length. (Adapted from International Human Genome Sequencing Consortium, *Nature* 409:860–921, 2001. With permission from Macmillan Publishers Ltd.)

marks the downstream exon boundary, and U2AF, which specifies the upstream one. By specifically marking the exons in this way and thereby taking advantage of the relatively uniform size of exons, the cell increases the accuracy with which it deposits the initial splicing components on the nascent RNA and thereby helps to avoid cryptic splice sites. How the SR proteins discriminate exon sequences from intron sequences is not understood in detail; however, it is known that some of the SR proteins bind preferentially to specific RNA sequences in exons, termed *splicing enhancers*. In principle, since any one of several different codons can be used to code for a given amino acid, there is freedom to adjust the exon nucleotide sequence so as to form a binding site for an SR protein, without necessarily affecting the amino acid sequence that the exon specifies.

Both the marking of exon and intron boundaries and the assembly of the spliceosome begin on an RNA molecule while it is still being elongated by RNA polymerase at its 3′ end. However, the actual chemistry of splicing can take place much later. This delay means that intron sequences are not necessarily removed from a pre-mRNA molecule in the order in which they occur along the RNA chain. It also means that, although spliceosome assembly is co-transcriptional, the splicing reactions sometimes occur posttranscriptionally—that is, after a complete pre-mRNA molecule has been made.

## A Second Set of snRNPs Splice a Small Fraction of Intron Sequences in Animals and Plants

Simple eucaryotes such as yeasts have only one set of snRNPs that perform all pre-mRNA splicing. However, more complex eucaryotes such as flies, mammals, and plants have a second set of snRNPs that direct the splicing of a small fraction of their intron sequences. This minor form of spliceosome recognizes a different set of RNA sequences at the 5′ and 3′ splice junctions and at the branch point; it is called the *U12-type spliceosome* because of the involvement of the

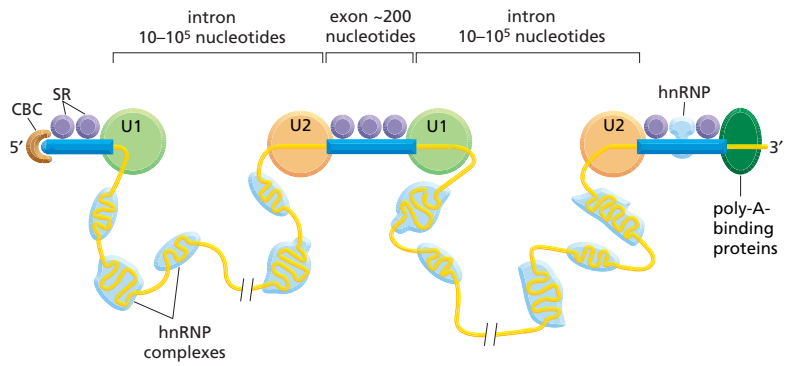



**Figure 6–33 The exon definition idea.** According to one proposal, SR proteins bind to each exon sequence in the pre-mRNA and thereby help to guide the snRNPs to the proper intron/exon boundaries. This demarcation of exons by the SR proteins occurs co-transcriptionally, beginning at the CBC (cap-binding complex) at the 5′ end. As indicated, the intron sequences in the pre-mRNA, which can be extremely long, are packaged into hnRNP (heterogeneous nuclear ribonucleoprotein) complexes that compact them into more manageable structures and perhaps mask cryptic splice sites. It has been proposed that hnRNP proteins may preferentially associate with intron sequences and that this preference may also help the spliceosome distinguish introns from exons. However, as shown, at least some hnRNP proteins also bind to exon sequences. (Adapted from R. Reed, *Curr. Opin. Cell Biol.* 12:340–345, 2000. With permission from Elsevier.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 100 of 240

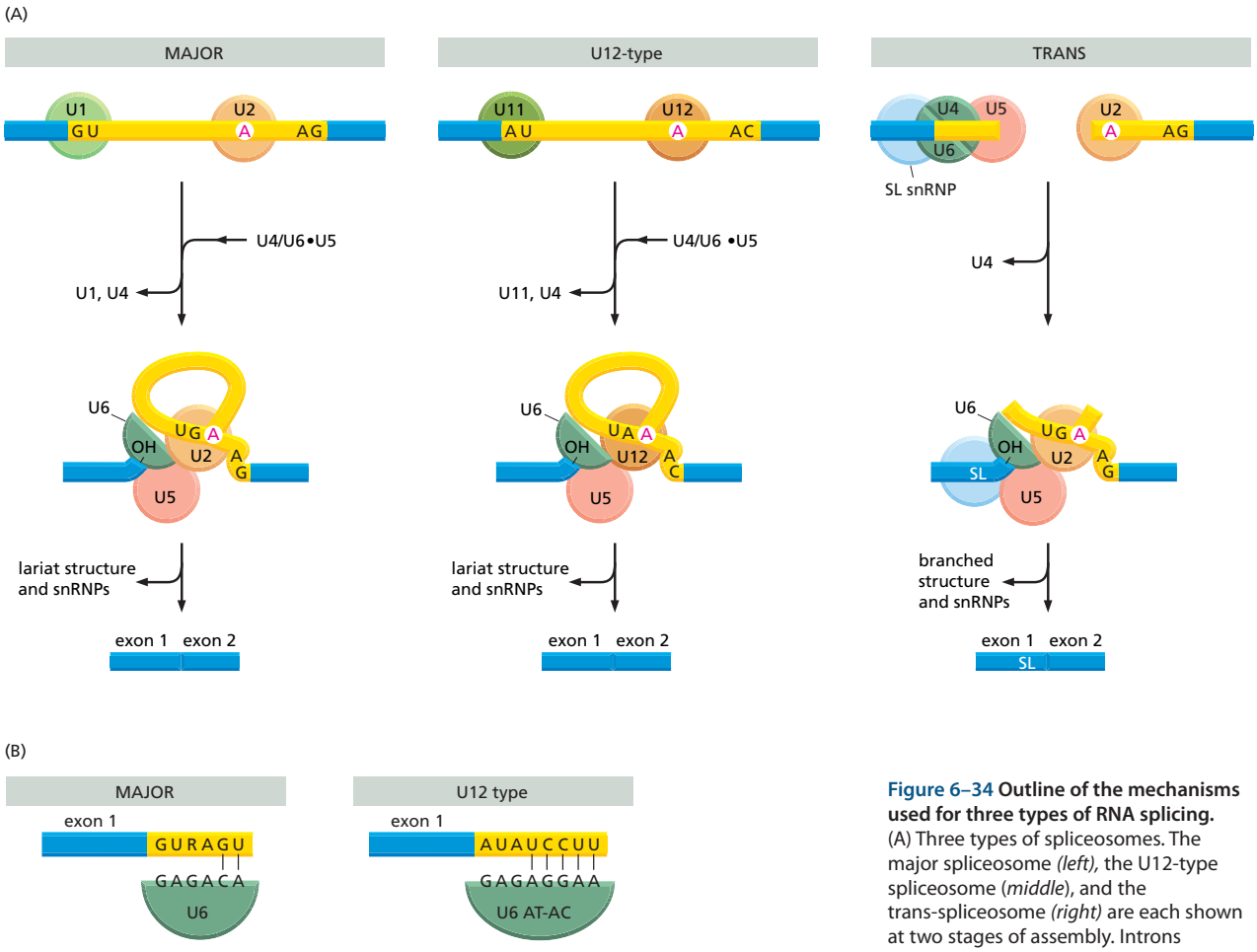



**Figure 6–34 Outline of the mechanisms used for three types of RNA splicing.** (A) Three types of spliceosomes. The major spliceosome *(left)*, the U12-type spliceosome (*middle*), and the trans-spliceosome *(right)* are each shown at two stages of assembly. Introns removed by the U12-type spliceosome have a different set of consensus nucleotide sequences from those removed by the major spliceosome. In humans, it is estimated that 0.1% of introns are removed by the U12-type spliceosome. In trans-splicing, which does not occur in humans, the SL snRNP is consumed in the reaction because a portion of the SL snRNA becomes the first exon of the mature mRNA. (B) The major U6 snRNP and the U6 snRNP specific to the U12-type spliceosome both recognize the 5′ splice junction, but they do so through a different set of base-pair interactions. The sequences shown are from humans. (Adapted from Y.T. Yu et al., The RNA World, pp. 487–524. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press, 1999.)

U12 SnRNP (**Figure 6–34**A). Despite recognizing different nucleotide sequences, the snRNPs in this spliceosome make the same types of RNA–RNA interactions with the pre-mRNA and with each other as do the major snRNPs (Figure 6–34B). Although, as we have seen, components of the major spliceosomes travel with RNA polymerase II as it transcribes genes, this may not be the case for the U12 spliceosome. It is possible that U12-mediated splicing is thereby delayed, and this presents the cell with a way to co-regulate splicing of the several hundred genes whose expression requires this spliceosome. A number of mammalian mRNAs contain a mixture of introns, some removed by the major spliceosome and others by the minor spliceosome, and it has been proposed that this arrangement permits particularly complex patterns of alternative splicing to occur.

A few eucaryotic organisms exhibit a particular variation on splicing, called **trans-splicing**. These organisms include the single-celled trypanosomes—protozoans that cause African sleeping sickness in humans—and the model multicellular organism, the nematode worm. In trans-splicing, exons from two separate RNA transcripts are spliced together to form a mature mRNA molecule (see Figure 6–34A). Trypanosomes produce all of their mRNAs in this way, whereas trans-splicing accounts for only about 1% of nematode mRNAs. In both cases, a single exon is spliced onto the 5′ end of many different RNA transcripts produced by the cell; in this way, all of the products of trans-splicing have the same 5′ exon and different 3′ exons. Many of the same snRNPs that function in conventional splicing are used in this reaction, although trans-splicing uses a unique snRNP (called the SL RNP) that brings in the common exon (see Figure 6–34).

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 101 of 240

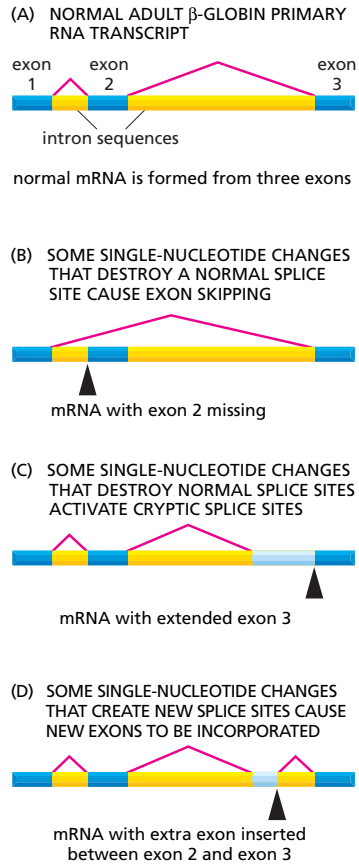



**Figure 6–35 Abnormal processing of the β-globin primary RNA transcript in humans with the disease β thalassemia.** In the examples shown, the disease is caused by splice-site mutations (*black arrowheads*) found in the genomes of affected patients. The *dark blue boxes* represent the three normal exon sequences; the *red lines* indicate the 5′ and 3′ splice sites. The *light blue boxes* depict new nucleotide sequences included in the final mRNA molecule as a result of the mutation. Note that when a mutation leaves a normal splice site without a partner, an exon is skipped or one or more abnormal cryptic splice sites nearby is used as the partner site, as in (C) and (D). (Adapted in part from S.H. Orkin, in The Molecular Basis of Blood Diseases [G. Stamatoyannopoulos et al., eds.], pp. 106–126. Philadelphia: Saunders, 1987.)

We do not know why even a few organisms use trans-splicing; however, it is thought that the common 5′ exon may aid in the translation of the mRNA. Thus, the mRNAs produced by trans-splicing in nematodes seem to be translated with especially high efficiency.

## RNA Splicing Shows Remarkable Plasticity

We have seen that the choice of splice sites depends on such features of the pre-mRNA transcript as the affinity of the three signals on the RNA (the 5′ and 3′ splice junctions and the branch point) for the splicing machinery, the co-transcriptional assembly of the spliceosome, and the "bookkeeping" that underlies exon definition. We do not know how accurate splicing normally is because, as we see later, there are several quality control systems that rapidly destroy mRNAs whose splicing goes awry. However, we do know that, compared with other steps in gene expression, splicing is unusually flexible. For example, a mutation in a nucleotide sequence critical for splicing of a particular intron does not necessarily prevent splicing of that intron altogether. Instead, the mutation typically creates a new pattern of splicing (**Figure 6–35**). Most commonly, an exon is simply skipped (Figure 6–35B). In other cases, the mutation causes a cryptic splice junction to be efficiently used (Figure 6–35C). Apparently, the splicing machinery has evolved to pick out the best possible pattern of splice junctions, and if the optimal one is damaged by mutation, it will seek out the next best pattern, and so on. This flexibility in the process of RNA splicing suggests that changes in splicing patterns caused by random mutations have been an important pathway in the evolution of genes and organisms.

The plasticity of RNA splicing also means that the cell can regulate the pattern of RNA splicing. Earlier in this section we saw that alternative splicing can give rise to different proteins from the same gene. Some examples of alternative splicing are constitutive; that is, the alternatively spliced mRNAs are produced continuously by cells of an organism. However, in many cases, the cell regulates the splicing patterns so that different forms of the protein are produced at different times and in different tissues (see Figure 6–27). In Chapter 7 we return to this issue to discuss some specific examples of regulated RNA splicing.

## Spliceosome-Catalyzed RNA Splicing Probably Evolved from Self-splicing Mechanisms

When the spliceosome was first discovered, it puzzled molecular biologists. Why do RNA molecules instead of proteins perform important roles in splice site recognition and in the chemistry of splicing? Why is a lariat intermediate used rather than the apparently simpler alternative of bringing the 5′ and 3′ splice sites together in a single step, followed by their direct cleavage and rejoining? The answers to these questions reflect the way in which the spliceosome is believed to have evolved.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 102 of 240

As discussed briefly in Chapter 1 (and in more detail in the final section of this chapter), it is likely that early cells used RNA molecules rather than proteins as their major catalysts and that they stored their genetic information in RNA rather than in DNA sequences. RNA-catalyzed splicing reactions presumably had important roles in these early cells. As evidence, some *self-splicing RNA* introns (that is, intron sequences in RNA whose splicing out can occur in the absence of proteins or any other RNA molecules) remain today—for example, in the nuclear rRNA genes of the ciliate *Tetrahymena*, in a few bacteriophage T4 genes, and in some mitochondrial and chloroplast genes.

A self-splicing intron sequence can be identified in a test tube by incubating a pure RNA molecule that contains the intron sequence and observing the splicing reaction. Two major classes of self-splicing intron sequences can be distinguished in this way. *Group I intron sequences* begin the splicing reaction by binding a G nucleotide to the intron sequence; this G is thereby activated to form the attacking group that will break the first of the phosphodiester bonds cleaved during splicing (the bond at the 5′ splice site). In *group II intron sequences,* an especially reactive A residue in the intron sequence is the attacking group, and a lariat intermediate is generated. Otherwise the reaction pathways for the two types of self-splicing intron sequences are the same. Both are presumed to represent vestiges of very ancient mechanisms (**Figure 6–36**).

For both types of self-splicing reactions, the nucleotide sequence of the intron is critical; the intron RNA folds into a specific three-dimensional structure, which brings the 5′ and 3′ splice junctions together and provides precisely positioned reactive groups to perform the chemistry (see Figure 6–6C). Because the chemistries of their splicing reactions are so similar, it has been proposed that the pre-mRNA splicing mechanism of the spliceosome evolved from group

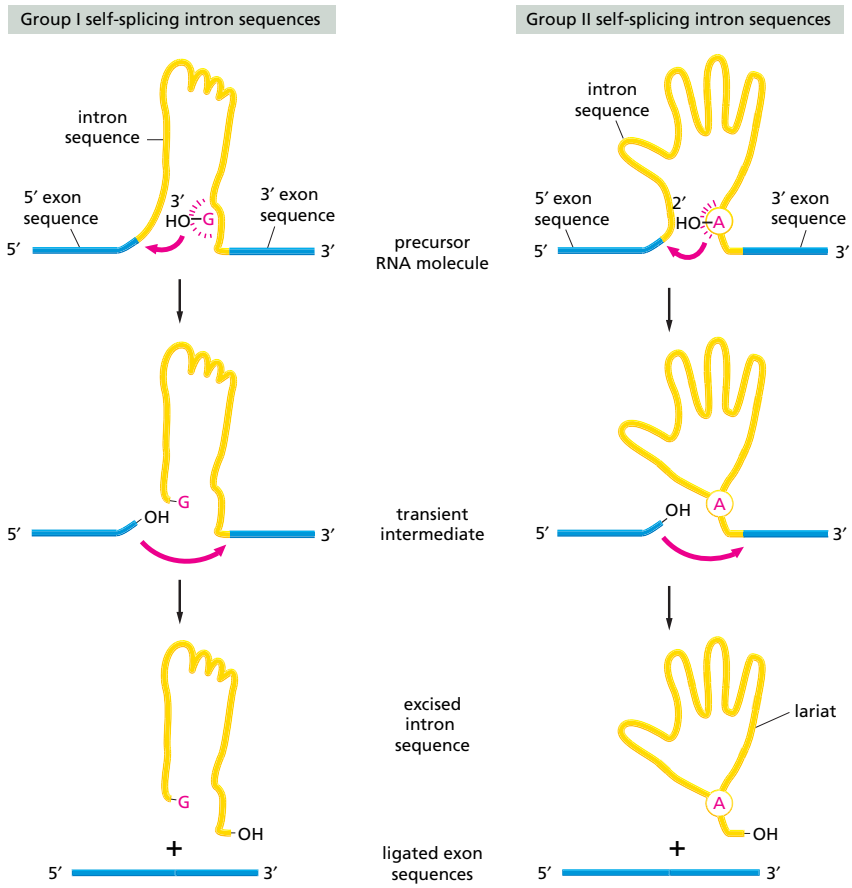



**Figure 6–36 The two known classes of self-splicing intron sequences.** The figure emphasizes the similarities between the two mechanisms. Both are normally aided by proteins in the cell that speed up the reaction, but the catalysis is nevertheless mediated by the RNA in the intron sequence. The group I intron sequences bind a free G nucleotide to a specific site on the RNA to initiate splicing, while the group II intron sequences use an especially reactive A nucleotide in the intron sequence itself for the same purpose. Both types of self-splicing reactions require the intron to be folded into a highly specific three-dimensional structure that provides the catalytic activity for the reaction (see Figure 6–6). The mechanism used by group II intron sequences releases the intron as a lariat structure and closely resembles the pathway of pre-mRNA splicing catalyzed by the spliceosome (compare with Figure 6–29). The spliceosome performs most RNA splicing in eucaryotic cells, and self-splicing RNAs represent unusual cases. (Adapted from T.R. Cech, *Cell* 44:207–210, 1986. With permission from Elsevier.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 103 of 240

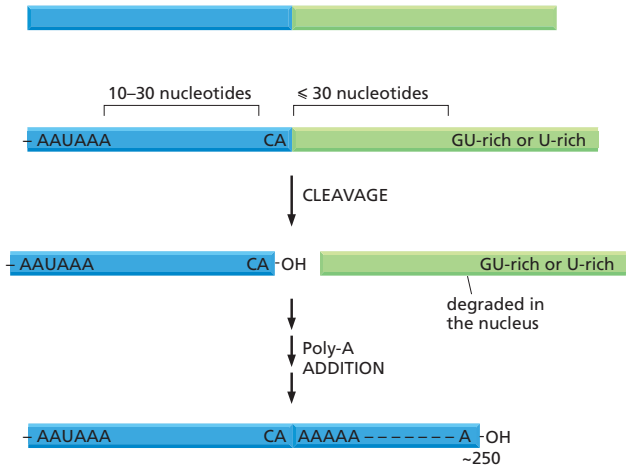



**Figure 6–37 Consensus nucleotide sequences that direct cleavage and polyadenylation to form the 3′ end of a eucaryotic mRNA.** These sequences are encoded in the genome; specific proteins recognize them after they are transcribed into RNA. The hexamer AAUAAA is bound by CPSF, the GU-rich element beyond the cleavage site is bound by CstF (see Figure 6–38), and the CA sequence is bound by a third factor required for the cleavage step. Like other consensus nucleotide sequences discussed in this chapter (see Figure 6–12), the sequences shown in the figure represent a variety of individual cleavage and polyadenylation signals.

II self-splicing. According to this idea, when the spliceosomal snRNPs took over the structural and chemical roles of the group II introns, the strict sequence constraints on intron sequences would have disappeared, thereby permitting a vast expansion in the number of different RNAs that could be spliced.

## RNA-Processing Enzymes Generate the 3′ End of Eucaryotic mRNAs

As previously explained, the 5′ end of the pre-mRNA produced by RNA polymerase II is capped almost as soon as it emerges from the RNA polymerase. Then, as the polymerase continues its movement along a gene, the spliceosome assembles on the RNA and delineates the intron and exon boundaries. The long C-terminal tail of the RNA polymerase coordinates these processes by transferring capping and splicing components directly to the RNA as it emerges from the enzyme. We see in this section that, as RNA polymerase II reaches the end of a gene, a similar mechanism ensures that the 3′ end of the pre-mRNA is appropriately processed.

As might be expected, the position of the 3′ end of each mRNA molecule is ultimately specified by a signal encoded in the genome (**Figure 6–37**). These signals are transcribed into RNA as the RNA polymerase II moves through them, and they are then recognized (as RNA) by a series of RNA-binding proteins and RNA-processing enzymes (**Figure 6–38**). Two multisubunit proteins, called CstF (cleavage stimulation factor) and CPSF (cleavage and polyadenylation specificity factor), are of special importance. Both of these proteins travel with the RNA polymerase tail and are transferred to the 3′-end processing sequence on an RNA molecule as it emerges from the RNA polymerase.

Once CstF and CPSF bind to specific nucleotide sequences on the emerging RNA molecule, additional proteins assemble with them to create the 3′ end of the mRNA. First, the RNA is cleaved (see Figure 6–38). Next an enzyme called poly-A polymerase (PAP) adds, one at a time, approximately 200 A nucleotides to the 3′ end produced by the cleavage. The nucleotide precursor for these additions is ATP, and the same type of 5′-to-3′ bonds are formed as in conventional RNA synthesis (see Figure 6–4). Unlike the usual RNA polymerases, poly-A polymerase does not require a template; hence the poly-A tail of eucaryotic mRNAs

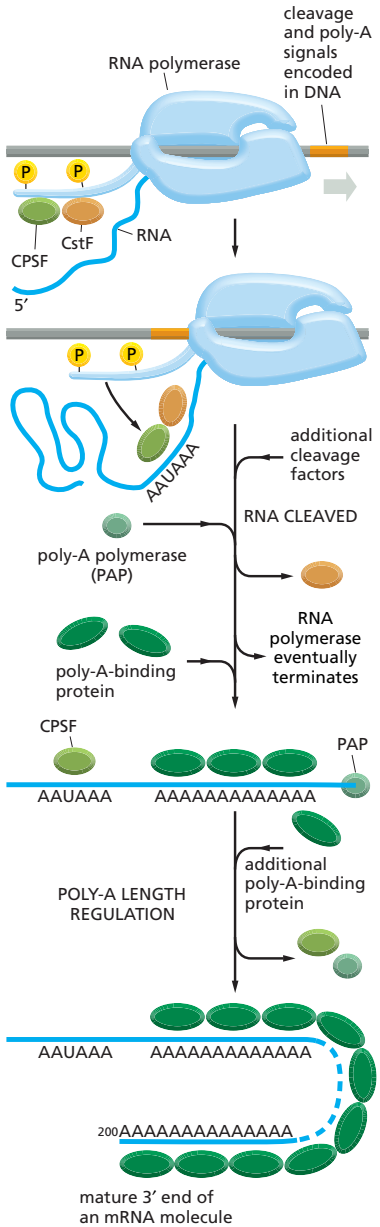



**Figure 6–38 Some of the major steps in generating the 3′ end of a eucaryotic mRNA.** This process is much more complicated than the analogous process in bacteria, where the RNA polymerase simply stops at a termination signal and releases both the 3′ end of its transcript and the DNA template (see Figure 6–11).

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 104 of 240

is not directly encoded in the genome. As the poly-A tail is synthesized, proteins called poly-A-binding proteins assemble onto it and, by a poorly understood mechanism, determine the final length of the tail. Some poly-A-binding proteins remain bound to the poly-A tail as the mRNA travels from the nucleus to the cytosol and they help to direct the synthesis of a protein on the ribosome, as we see later in this chapter.

After the 3′ end of a eucaryotic pre-mRNA molecule has been cleaved, the RNA polymerase II continues to transcribe, in some cases for hundreds of nucleotides. But the polymerase soon releases its grip on the template and transcription terminates. After 3′-end cleavage has occurred, the newly synthesized RNA that emerges from the polymerases lacks a 5′ cap; this unprotected RNA is rapidly degraded by a 5′ → 3′ exonuclease, which is carried along on the polymerase tail. Apparently, it is this RNA degradation that eventually causes the RNA polymerase to dissociate from the DNA.

## Mature Eucaryotic mRNAs Are Selectively Exported from the Nucleus

We have seen how eucaryotic pre-mRNA synthesis and processing take place in an orderly fashion within the cell nucleus. However, these events create a special problem for eucaryotic cells, especially those of complex organisms where the introns are vastly longer than the exons. Of the pre-mRNA that is synthesized, only a small fraction—the mature mRNA—is of further use to the cell. The rest—excised introns, broken RNAs, and aberrantly processed pre-mRNAs—is not only useless but potentially dangerous. How, then, does the cell distinguish between the relatively rare mature mRNA molecules it wishes to keep and the overwhelming amount of debris from RNA processing?

The answer is that, as an RNA molecule is processed, it loses certain proteins and acquires others, thereby signifying the successful completion of each of the different steps. For example, we have seen that acquisition of the cap-binding complexes, the exon junction complexes, and the poly-A-binding proteins mark the completion of capping, splicing, and poly-A addition, respectively. A properly completed mRNA molecule is also distinguished by the proteins it lacks. For example, the presence of a snRNP would signify incomplete or aberrant splicing. Only when the proteins present on an mRNA molecule collectively signify that processing was successfully completed is the mRNA exported from the nucleus into the cytosol, where it can be translated into protein. Improperly processed mRNAs, and other RNA debris are retained in the nucleus, where they are eventually degraded by the nuclear **exosome**, a large protein complex whose interior is rich in 3′-to-5′ RNA exonucleases. Eucaryotic cells thus export only useful RNA molecules to the cytoplasm, while debris is disposed of in the nucleus.

Of all the proteins that assemble on pre-mRNA molecules as they emerge from transcribing RNA polymerases, the most abundant are the hnRNPs (heterogeneous nuclear ribonuclear proteins) (see Figure 6–33). Some of these proteins (there are approximately 30 of them in humans) unwind the hairpin helices from the RNA so that splicing and other signals on the RNA can be read more easily. Others preferentially package the RNA contained in the very long intron sequences typically found in genes of complex organisms. They may therefore play an important role in distinguishing mature mRNA from the debris left over from RNA processing.

Successfully processed mRNAs are guided through the **nuclear pore complexes** (NPCs)—aqueous channels in the nuclear membrane that directly connect the nucleoplasm and cytosol (**Figure 6–39**). Small molecules (less than 50,000 daltons) can diffuse freely through these channels. However, most of the macromolecules in cells, including mRNAs complexed with proteins, are far too large to pass through the channels without a special process. The cell uses energy to actively transport such macromolecules in both directions through the nuclear pore complexes.

As explained in detail in Chapter 12, macromolecules are moved through nuclear pore complexes by *nuclear transport receptors,* which, depending on the

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 105 of 240

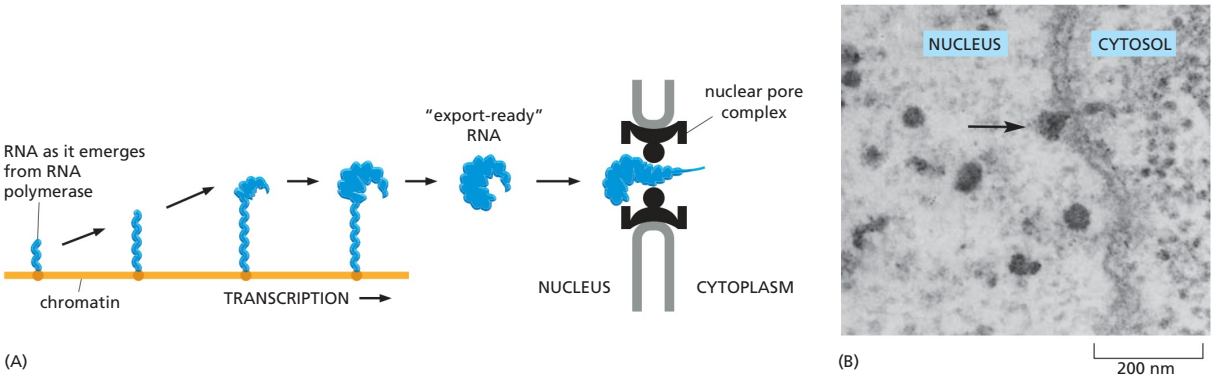



**Figure 6–39 Transport of a large mRNA molecule through the nuclear pore complex.** (A) The maturation of an mRNA molecule as it is synthesized by RNA polymerase and packaged by a variety of nuclear proteins. This drawing of an unusually abundant RNA, called the Balbiani Ring mRNA, is based on EM micrographs such as that shown in (B). Balbiani Rings are found in the cells of certain insects. (A, adapted from B. Daneholt, *Cell* 88:585–588, 1997. With permission from Elsevier; B, from B.J. Stevens and H. Swift, *J. Cell Biol*. 31:55–77, 1966. With permission from The Rockefeller University Press.)

identity of the macromolecule, escort it from the nucleus to the cytoplasm or vice versa. For mRNA export to occur, a specific nuclear transport receptor must be loaded onto the mRNA, a step that, at least in some organisms, takes place in concert with 3′ cleavage and polyadenylation. Once it helps to move an RNA molecule through the nuclear pore complex, the transport receptor dissociates from the mRNA, re-enters the nucleus, and exports a new mRNA molecule (**Figure 6–40**).

The export of mRNA–protein complexes from the nucleus can be observed with the electron microscope for the unusually abundant mRNA of the insect Balbiani Ring genes. As these genes are transcribed, the newly formed RNA is seen to be packaged by proteins, including hnRNPs, SR proteins, and components of the spliceosome. This protein–RNA complex undergoes a series of structural transitions, probably reflecting RNA processing events, culminating in a curved fiber (see Figure 6–39). This curved fiber moves through the nucleoplasm and enters the nuclear pore complex (with its 5′ cap proceeding first), and it then undergoes another series of structural transitions as it moves through the

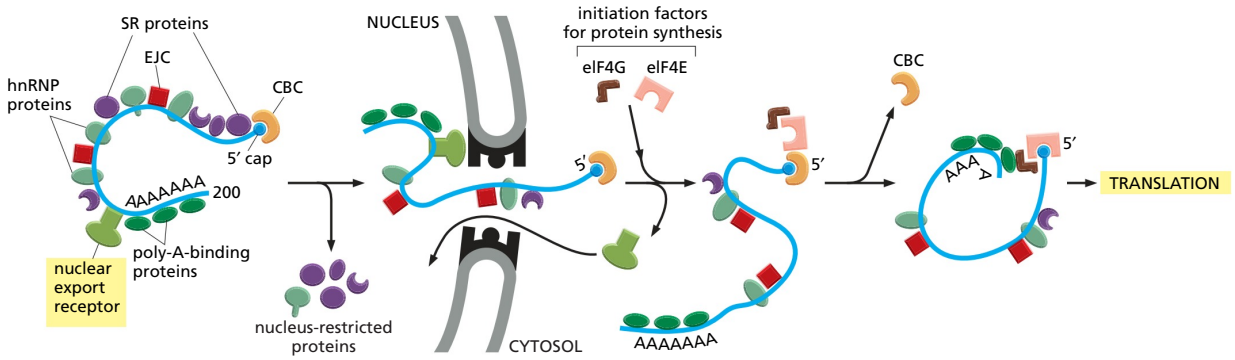



**Figure 6–40 Schematic illustration of an "export-ready" mRNA molecule and its transport through the nuclear pore.** As indicated, some proteins travel with the mRNA as it moves through the pore, whereas others remain in the nucleus. The nuclear export receptor for mRNAs is a complex of proteins that is deposited when the mRNA has been correctly spliced and polyadenylated. When the mRNA is exported to the cytosol, the export receptor dissociates from the mRNA and is re-imported into the nucleus, where it can be used again. Just after it leaves the nucleus, and before it loses the cap-binding complex (CBC) the mRNA is subjected to a final check, called *nonsense-mediated decay*, which is described later in the chapter. Once it passes this test the mRNA continues to shed previously bound proteins and acquire new ones before it is efficiently translated into protein. EJC, exon junction complex.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 106 of 240

pore. These and other observations reveal that the pre-mRNA–protein and mRNA–protein complexes are dynamic structures that gain and lose numerous specific proteins during RNA synthesis, processing, and export (see Figure 6–40).

As we have seen, some of these proteins mark the different stages of mRNA maturation; other proteins deposited on the mRNA while it is still in the nucleus can affect the fate of the RNA after it is transported to the cytosol. Thus, the stability of an mRNA in the cytosol, the efficiency with which it is translated into protein, and its ultimate destination in the cytosol can all be determined by proteins acquired in the nucleus that remain bound to the RNA after it leaves the nucleus. We will discuss these issues in Chapter 7 when we turn to the post-transcriptional control of gene expression.

We have seen that RNA synthesis and processing are closely coupled in the cell, and it might be expected that export from the nucleus is somehow integrated with these two processes. Although the Balbiani Ring RNAs can be seen to move through the nucleoplasm and out through the nuclear pores, other mRNAs appear to be synthesized and processed in close proximity to nuclear pore complexes. In these cases, which may represent the majority of eucaryotic genes, mRNA synthesis, processing, and transport all appear to be tightly coupled; the mRNA can thus be viewed as emerging from the nuclear pore as a newly manufactured car might emerge from an assembly line. Later in this chapter, we will see that the cell performs an additional quality-control check on each mRNA before it is allowed to be efficiently translated into protein.

Before discussing what happens to mRNAs after they leave the nucleus, we briefly consider how the synthesis and processing of noncoding RNA molecules occurs. Although there are many other examples, our discussion focuses on the rRNAs that are critically important for the translation of mRNAs into protein.

## Many Noncoding RNAs Are Also Synthesized and Processed in the Nucleus

A few percent of the dry weight of a mammalian cell is RNA; of that, only about 3–5% is mRNA. A fraction of the remainder represents intron sequences before they have been degraded, but the bulk of the RNA in cells performs structural and catalytic functions (see Table 6–1, p. 336). The most abundant RNAs in cells are the ribosomal RNAs (rRNAs), constituting approximately 80% of the RNA in rapidly dividing cells. As discussed later in this chapter, these RNAs form the core of the ribosome. Unlike bacteria—in which a single RNA polymerase synthesizes all RNAs in the cell—eucaryotes have a separate, specialized polymerase, RNA polymerase I, that is dedicated to producing rRNAs. RNA polymerase I is similar structurally to the RNA polymerase II discussed previously; however, the absence of a C-terminal tail in polymerase I helps to explain why its transcripts are neither capped nor polyadenylated. As mentioned earlier, this difference helps the cell distinguish between noncoding RNAs and mRNAs.

Because multiple rounds of translation of each mRNA molecule can provide an enormous amplification in the production of protein molecules, many of the proteins that are very abundant in a cell can be synthesized from genes that are present in a single copy per haploid genome. In contrast, the RNA components of the ribosome are final gene products, and a growing mammalian cell must synthesize approximately 10 million copies of each type of ribosomal RNA in each cell generation to construct its 10 million ribosomes. The cell can produce adequate quantities of ribosomal RNAs only because it contains multiple copies of the **rRNA genes** that code for ribosomal RNAs (**rRNAs**). Even *E. coli* needs seven copies of its rRNA genes to meet the cell's need for ribosomes. Human cells contain about 200 rRNA gene copies per haploid genome, spread out in small clusters on five different chromosomes (see Figure 4–11), while cells of the frog *Xenopus* contain about 600 rRNA gene copies per haploid genome in a single cluster on one chromosome (**Figure 6–41**).

There are four types of eucaryotic rRNAs, each present in one copy per ribosome. Three of the four rRNAs (18S, 5.8S, and 28S) are made by chemically modifying and cleaving a single large precursor rRNA (**Figure 6–42**); the fourth (5S

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 107 of 240

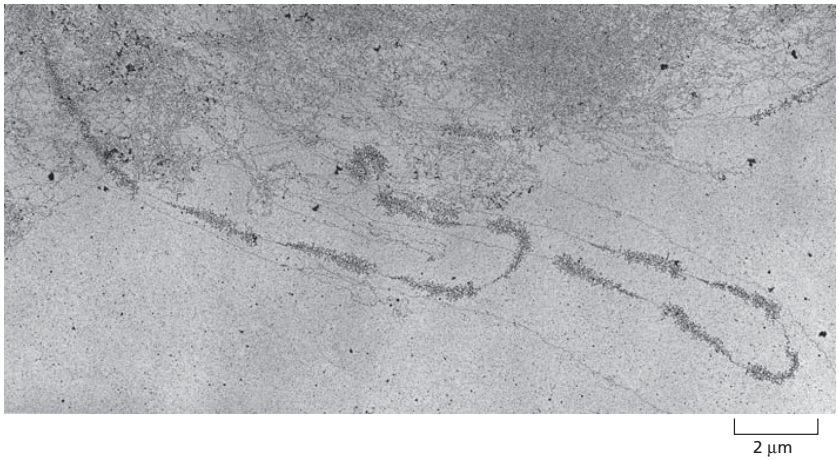



**Figure 6–41 Transcription from tandemly arranged rRNA genes, as seen in the electron microscope.** The pattern of alternating transcribed gene and nontranscribed spacer is readily seen. A higher-magnification view of rRNA genes is shown in Figure 6–9. (From V.E. Foe, *Cold Spring Harbor Symp. Quant. Biol*. 42:723–740, 1978. With permission from Cold Spring Harbor Laboratory Press.)

RNA) is synthesized from a separate cluster of genes by a different polymerase, RNA polymerase III, and does not require chemical modification.

Extensive chemical modifications occur in the 13,000-nucleotide-long precursor rRNA before the rRNAs are cleaved out of it and assembled into ribosomes. These include about 100 methylations of the 2′-OH positions on nucleotide sugars and 100 isomerizations of uridine nucleotides to pseudouridine (**Figure 6–43**A). The functions of these modifications are not understood in detail, but many probably aid in the folding and assembly of the final rRNAs and some may subtly alter the function of ribosomes. Each modification is made at a specific position in the precursor rRNA. These positions are specified by about 150 "guide RNAs," which position themselves through base-pairing to the precursor rRNA and thereby bring an RNA-modifying enzyme to the appropriate position (Figure 6–43B). Other guide RNAs promote cleavage of the precursor rRNAs into the mature rRNAs, probably by causing conformational changes in the precursor rRNA that expose these sites to nucleases. All of these guide RNAs are members of a large class of RNAs called **small nucleolar RNAs** (or **snoRNAs**), so named because these RNAs perform their functions in a subcompartment of the nucleus called the nucleolus. Many snoRNAs are encoded in the introns of other genes, especially those encoding ribosomal proteins. They are therefore synthesized by RNA polymerase II and processed from excised intron sequences.

Recently several snoRNA-like RNAs have been identified that are synthesized only in cells of the brain. These are believed to direct the modification of mRNAs, instead of rRNAs, and are likely to represent a new, but poorly understood, type of gene regulatory mechanism.

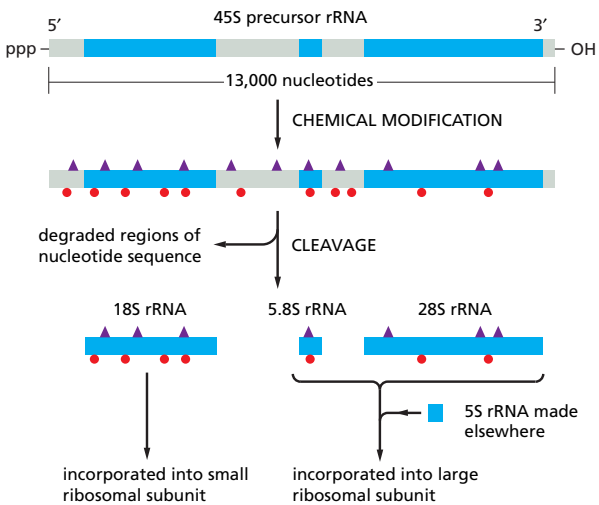



**Figure 6–42 The chemical modification and nucleolytic processing of a eucaryotic 45S precursor rRNA molecule into three separate ribosomal RNAs.** Two types of chemical modifications (color-coded as indicated in Figure 6–43) are made to the precursor rRNA before it is cleaved. Nearly half of the nucleotide sequences in this precursor rRNA are discarded and degraded in the nucleus. The rRNAs are named according to their "S" values, which refer to their rate of sedimentation in an ultracentrifuge. The larger the S value, the larger the rRNA.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 108 of 240

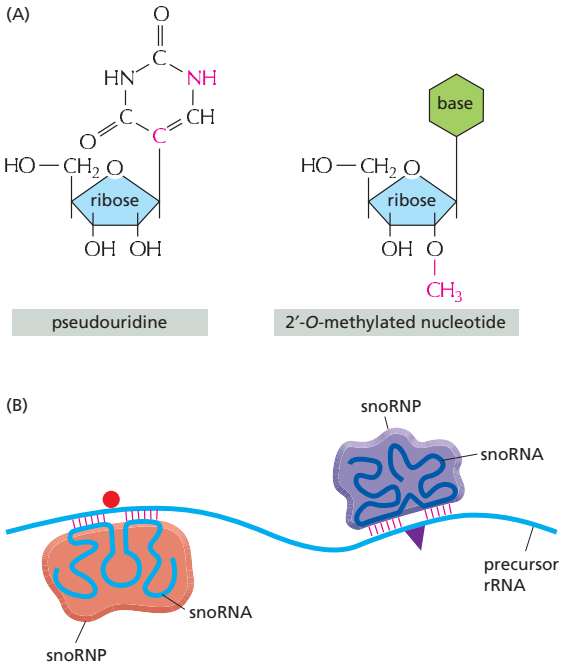



**Figure 6–43 Modifications of the precursor rRNA by guide RNAs.** (A) Two prominent covalent modifications occur after rRNA synthesis; the differences from the initially incorporated nucleotide are indicated by *red* atoms. Pseudouridine is an isomer of uridine; the base has been "rotated" relative to the sugar. (B) As indicated, snoRNAs determine the sites of modification by base-pairing to complementary sequences on the precursor rRNA. The snoRNAs are bound to proteins, and the complexes are called snoRNPs. snoRNPs contain both the guide sequences and the enzymes that modify the rRNA.

## The Nucleolus Is a Ribosome-Producing Factory

The nucleolus is the most obvious structure seen in the nucleus of a eucaryotic cell when viewed in the light microscope. Consequently, it was so closely scrutinized by early cytologists that an 1898 review could list some 700 references. We now know that the nucleolus is the site for the processing of rRNAs and their assembly into ribosome subunits. Unlike many of the major organelles in the cell, the nucleolus is not bound by a membrane (**Figure 6–44**); instead, it is a large aggregate of macromolecules, including the rRNA genes themselves, precursor rRNAs, mature rRNAs, rRNA-processing enzymes, snoRNPs, ribosomal proteins and partly assembled ribosomes. The close association of all these components presumably allows the assembly of ribosomes to occur rapidly and smoothly.

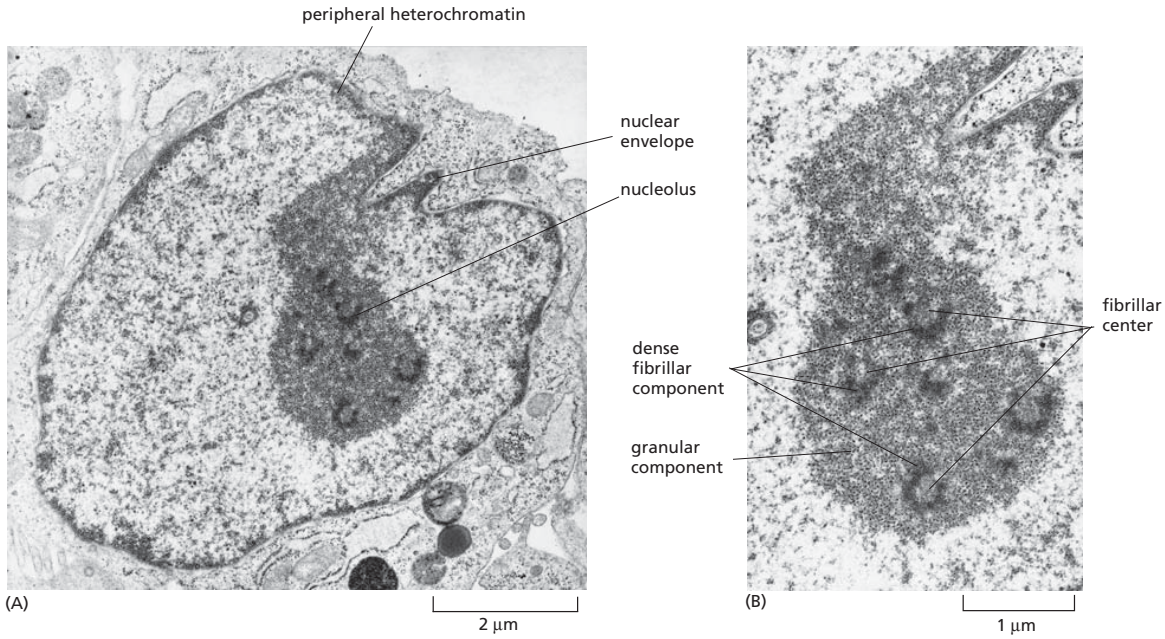



**Figure 6–44 Electron micrograph of a thin section of a nucleolus in a human fibroblast, showing its three distinct zones.** (A) View of entire nucleus. (B) High-power view of the nucleolus. It is believed that transcription of the rRNA genes takes place between the fibrillar center and the dense fibrillar component and that processing of the rRNAs and their assembly into the two subunits of the ribosome proceeds outward from the dense fibrillar component to the surrounding granular components. (Courtesy of E.G. Jordan and J. McGovern.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 109 of 240

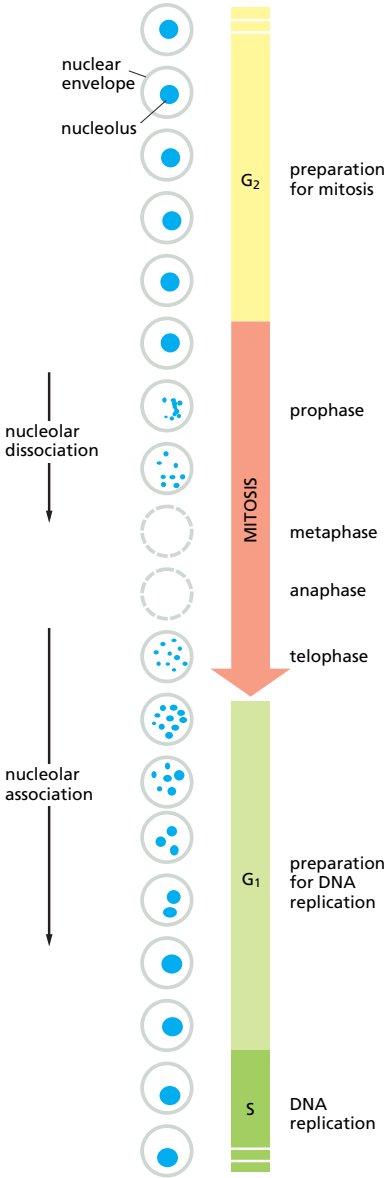



**Figure 6–45 Changes in the appearance of the nucleolus in a human cell during the cell cycle.** Only the cell nucleus is represented in this diagram. In most eucaryotic cells the nuclear envelope breaks down during mitosis, as indicated by the dashed circles.

Various types of RNA molecules play a central part in the chemistry and structure of the nucleolus, suggesting that it may have evolved from an ancient structure present in cells dominated by RNA catalysis. In present-day cells, the rRNA genes also have an important role in forming the nucleolus. In a diploid human cell, the rRNA genes are distributed into 10 clusters, located near the tips of five different chromosome pairs (see Figure 4–11). During interphase these 10 chromosomes contribute DNA loops (containing the rRNA genes) to the nucleolus; in M-phase, when the chromosomes condense, the nucleolus disappears. Finally, in the telophase part of mitosis, as chromosomes return to their semi-dispersed state, the tips of the 10 chromosomes coalesce and the nucleolus reforms (**Figure 6–45** and **Figure 6–46**). The transcription of the rRNA genes by RNA polymerase I is necessary for this process. As might be expected, the size of the nucleolus reflects the number of ribosomes that the cell is producing. Its size therefore varies greatly in different cells and can change in a single cell, occupying 25% of the total nuclear volume in cells that are making unusually large amounts of protein.

Ribosome assembly is a complex process, the most important features of which are outlined in **Figure 6–47**. In addition to its important role in ribosome biogenesis, the nucleolus is also the site where other RNAs are produced and other RNA–protein complexes are assembled. For example, the U6 snRNP, which functions in pre-mRNA splicing (see Figure 6–29), is composed of one RNA molecule and at least seven proteins. The U6 snRNA is chemically modified by snoRNAs in the nucleolus before its final assembly there into the U6 snRNP. Other important RNA–protein complexes, including telomerase (encountered in Chapter 5) and the signal recognition particle (which we discuss in Chapter 12), are also believed to be assembled at the nucleolus. Finally, the tRNAs (transfer RNAs) that carry the amino acids for protein synthesis are processed there as well; like the rRNA genes, those encoding tRNAs are clustered in the nucleolus. Thus, the nucleolus can be thought of as a large factory at which many different noncoding RNAs are transcribed, processed, and assembled with proteins to form a large variety of ribonucleoprotein complexes.

## The Nucleus Contains a Variety of Subnuclear Structures

Although the nucleolus is the most prominent structure in the nucleus, several other nuclear bodies have been observed and studied (**Figure 6–48**). These include Cajal bodies (named for the scientist who first described them in 1906), GEMS (Gemini of Cajal bodies), and interchromatin granule clusters (also called

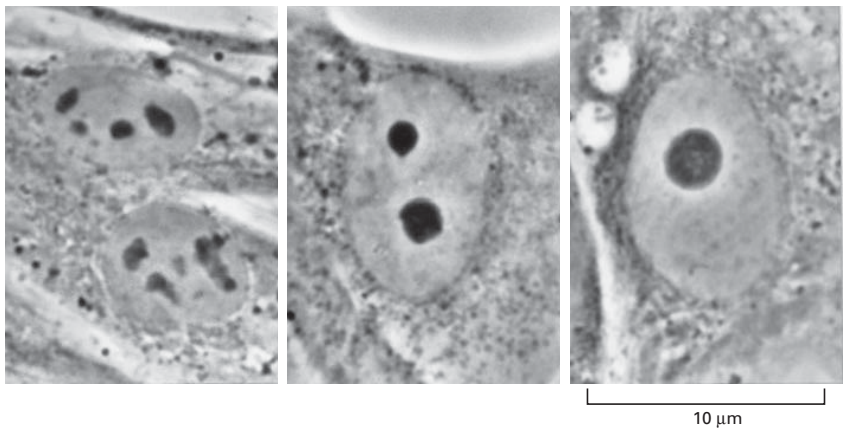



**Figure 6–46 Nucleolar fusion.** These light micrographs of human fibroblasts grown in culture show various stages of nucleolar fusion. After mitosis, each of the 10 human chromosomes that carry a cluster of rRNA genes begins to form a tiny nucleolus, but these rapidly coalesce as they grow to form the single large nucleolus typical of many interphase cells. (Courtesy of E.G. Jordan and J. McGovern.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 110 of 240

"speckles"). Like the nucleolus, these other nuclear structures lack membranes and are highly dynamic; their appearance is probably the result of the tight association of protein and RNA components involved in the synthesis, assembly, and storage of macromolecules involved in gene expression. Cajal bodies and GEMS resemble one another and are frequently paired in the nucleus; it is not clear whether they truly represent distinct structures. These are likely to be the locations in which snoRNAs and snRNAs undergo covalent modifications and final assembly with proteins. A group of guide RNAs, termed *small Cajal RNAs (scaRNAs),* selects the sites of these modifications through base pairing. Cajal bodies/GEMS may also be sites where the snRNPs are recycled and their RNAs are "reset" after the rearrangements that occur during splicing (see p. 352). In contrast, the interchromatin granule clusters have been proposed to be stockpiles of fully mature snRNPs and other RNA processing components that are ready to be used in the production of mRNA (**Figure 6–49**).

Scientists have had difficulties in working out the function of these small subnuclear structures, in part because their appearances differ between organisms and can change dramatically as cells traverse the cell cycle or respond to changes in their environment. Much of the progress now being made depends on genetic tools—examination of the effects of designed mutations in model

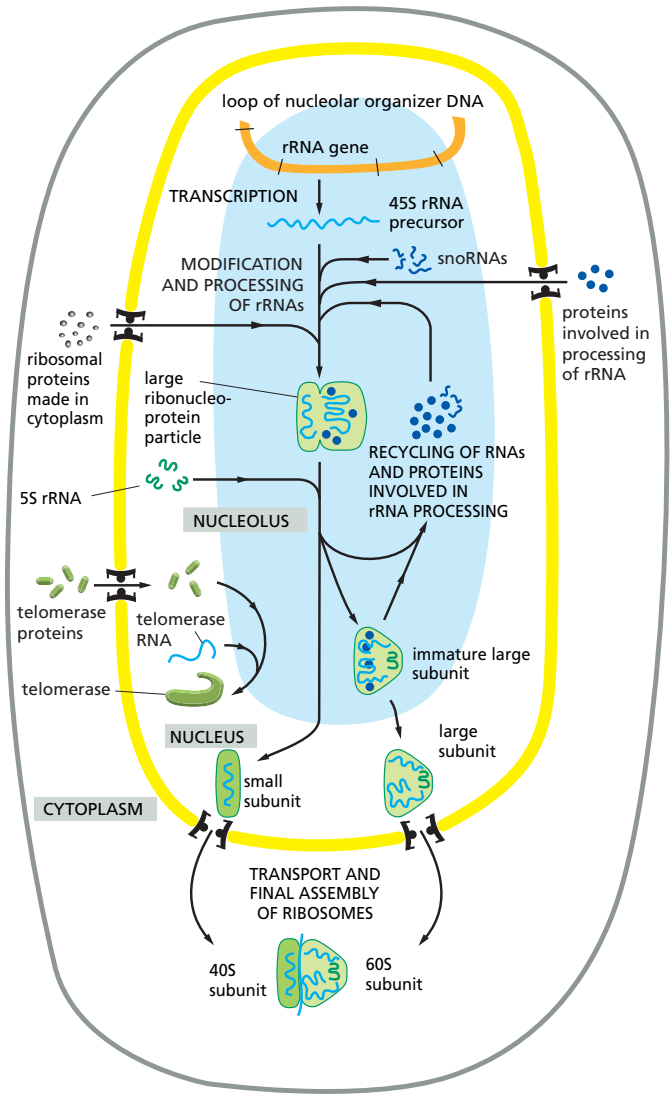



**Figure 6–47 The function of the nucleolus in ribosome and other ribonucleoprotein synthesis.** The 45S precursor rRNA is packaged in a large ribonucleoprotein particle containing many ribosomal proteins imported from the cytoplasm. While this particle remains at the nucleolus, selected pieces are added and others discarded as it is processed into immature large and small ribosomal subunits. The two ribosomal subunits are thought to attain their final functional form only as each is individually transported through the nuclear pores into the cytoplasm. Other ribonucleoprotein complexes, including telomerase shown here, are also assembled in the nucleolus.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 111 of 240

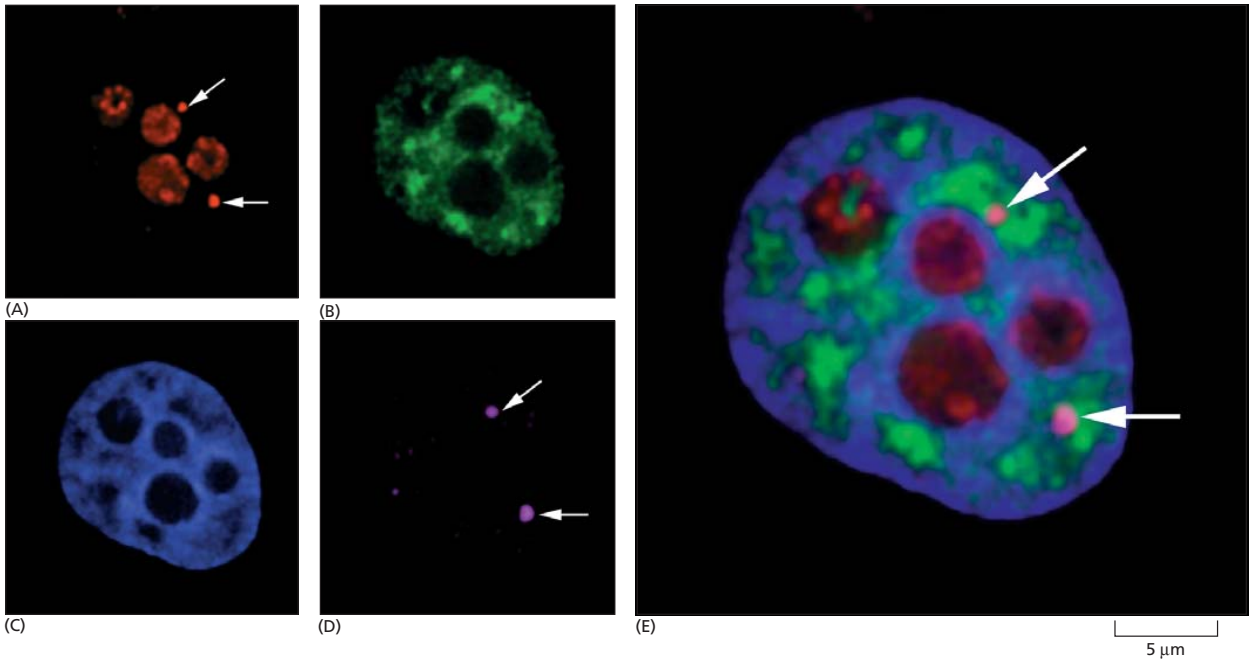



**Figure 6–48 Visualization of some prominent nuclear bodies.** (A)–(D) Micrographs of the same human cell nucleus, each processed to show a particular set of nuclear structures. (E) All four images enlarged and superimposed. (A) shows the location of the protein fibrillarin (a component of several snoRNPs), which is present at both nucleoli and Cajal bodies, the latter indicated by arrows. (B) shows interchromatin granule clusters or "speckles" detected by using antibodies against a protein involved in pre-mRNA splicing. (C) is stained to show bulk chromatin. (D) shows the location of the protein coilin, which is present at Cajal bodies (*arrows*; see also Figure 4–67). (From J.R. Swedlow and A.I. Lamond, *Gen. Biol.* 2:1–7, 2001. With permission from BioMed Central. Micrographs courtesy of Judith Sleeman.)

organisms or of spontaneous mutations in humans. As one example, GEMS contain the SMN (survival of motor neurons) protein. Certain mutations of the gene encoding this protein are the cause of inherited spinal muscular atrophy, a human disease characterized by a wasting away of the muscles. The disease seems to be caused by a defect in snRNP production. A more complete loss of snRNPs would be expected to be lethal.

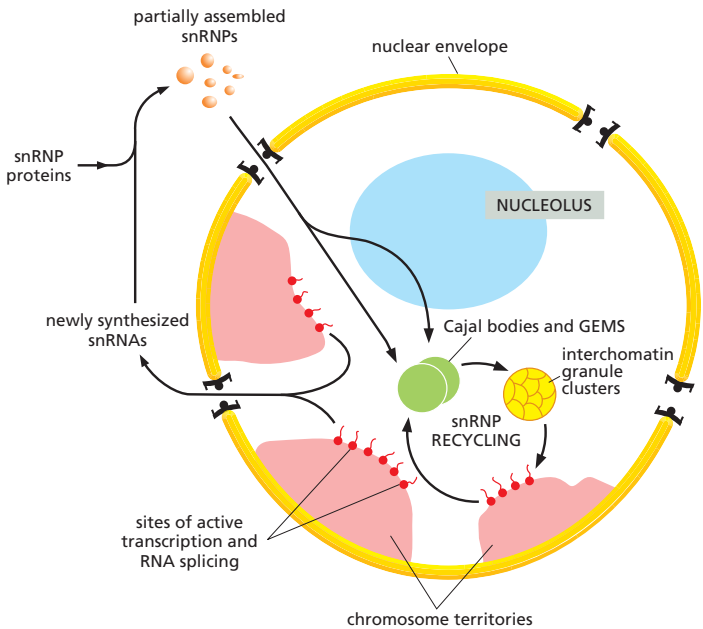



**Figure 6–49 Schematic view of subnuclear structures.** A typical vertebrate nucleus has several Cajal bodies, which are proposed to be the sites where snRNPs and snoRNPs undergo their final modifications. Interchromatin granule clusters are proposed to be storage sites for fully mature snRNPs. A typical vertebrate nucleus has 20–50 interchromatin granule clusters.

After their initial synthesis, snRNAs are exported from the nucleus to undergo 5′ and 3′ end-processing and assemble with the seven common snRNP proteins (called Sm proteins). These complexes are reimported into the nucleus and the snRNPs undergo their final modification by scaRNAs at Cajal bodies. In addition, snoRNAs chemically modify the U6 snRNP at the nucleolus. The sites of active transcription and splicing (approximately 2000–3000 sites per vertebrate nucleus) correspond to the "perichromatin fibers" seen under the electron microscope. (Adapted from J.D. Lewis and D. Tollervey, *Science* 288:1385–1389, 2000. With permission from AAAS.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 112 of 240

Given the importance of nuclear subdomains in RNA processing, it might have been expected that pre-mRNA splicing would occur in a particular location in the nucleus, as it requires numerous RNA and protein components. However, the assembly of splicing components on pre-mRNA is co-transcriptional; thus, splicing must occur at many locations along chromosomes. Although a typical mammalian cell may be expressing on the order of 15,000 genes, transcription and RNA splicing may be localized to only several thousand sites in the nucleus. These sites themselves are highly dynamic and probably result from the association of transcription and splicing components to create small "assembly lines" with a high local concentration of these components. Interchromatin granule clusters—which contain stockpiles of RNA-processing components—are often observed next to sites of transcription, as though poised to replenish supplies. Thus, the nucleus seems to be highly organized into subdomains, with snRNPs, snoRNPs, and other nuclear components moving between them in an orderly fashion according to the needs of the cell (see Figure 6–48; also see Figure 4–69).

### Summary

*Before the synthesis of a particular protein can begin, the corresponding mRNA molecule must be produced by transcription. Bacteria contain a single type of RNA polymerase (the enzyme that carries out the transcription of DNA into RNA). An mRNA molecule is produced when this enzyme initiates transcription at a promoter, synthesizes the RNA by chain elongation, stops transcription at a terminator, and releases both the DNA template and the completed mRNA molecule. In eucaryotic cells, the process of transcription is much more complex, and there are three RNA polymerases—polymerase I, II, and III—that are related evolutionarily to one another and to the bacterial polymerase.*

*RNA polymerase II synthesizes eucaryotic mRNA. This enzyme requires a series of additional proteins, the general transcription factors, to initiate transcription on a purified DNA template, and still more proteins (including chromatin-remodeling complexes and histone-modifying enzymes) to initiate transcription on its chromatin templates inside the cell.*

*During the elongation phase of transcription, the nascent RNA undergoes three types of processing events: a special nucleotide is added to its 5′ end (capping), intron sequences are removed from the middle of the RNA molecule (splicing), and the 3′ end of the RNA is generated (cleavage and polyadenylation). Each of these processes is initiated by proteins that travel along with RNA polymerase II by binding to sites on its long, extended C-terminal tail. Splicing is unusual in that many of its key steps are carried out by specialized RNA molecules rather than proteins. Properly processed mRNAs are passed through nuclear pore complexes into the cytosol, where they are translated into protein.*

*For some genes, RNA is the final product. In eucaryotes, these genes are usually transcribed by either RNA polymerase I or RNA polymerase III. RNA polymerase I makes the ribosomal RNAs. After their synthesis as a large precursor, the rRNAs are chemically modified, cleaved, and assembled into the two ribosomal subunits in the nucleolus—a distinct subnuclear structure that also helps to process some smaller RNA–protein complexes in the cell. Additional subnuclear structures (including Cajal bodies and interchromatin granule clusters) are sites where components involved in RNA processing are assembled, stored, and recycled.*

## FROM RNA TO PROTEIN

In the preceding section we have seen that the final product of some genes is an RNA molecule itself, such as those present in the snRNPs and in ribosomes. However, most genes in a cell produce mRNA molecules that serve as intermediaries on the pathway to proteins. In this section we examine how the cell converts the information carried in an mRNA molecule into a protein molecule. This feat of translation was a focus of attention of biologists in the late 1950s, when it

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 113 of 240

was posed as the "coding problem": how is the information in a linear sequence of nucleotides in RNA translated into the linear sequence of a chemically quite different set of units—the amino acids in proteins? This fascinating question stimulated great excitement among scientists at the time. Here was a cryptogram set up by nature that, after more than 3 billion years of evolution, could finally be solved by one of the products of evolution—human beings. And indeed, not only has the code been cracked step by step, but in the year 2000 the structure of the elaborate machinery by which cells read this code—the ribosome—was finally revealed in atomic detail.

## An mRNA Sequence Is Decoded in Sets of Three Nucleotides

Once an mRNA has been produced by transcription and processing, the information present in its nucleotide sequence is used to synthesize a protein. Transcription is simple to understand as a means of information transfer: since DNA and RNA are chemically and structurally similar, the DNA can act as a direct template for the synthesis of RNA by complementary base-pairing. As the term *transcription* signifies, it is as if a message written out by hand is being converted, say, into a typewritten text. The language itself and the form of the message do not change, and the symbols used are closely related.

In contrast, the conversion of the information in RNA into protein represents a **translation** of the information into another language that uses quite different symbols. Moreover, since there are only 4 different nucleotides in mRNA and 20 different types of amino acids in a protein, this translation cannot be accounted for by a direct one-to-one correspondence between a nucleotide in RNA and an amino acid in protein. The nucleotide sequence of a gene, through the intermediary of mRNA, is translated into the amino acid sequence of a protein by rules that are known as the **genetic code**. This code was deciphered in the early 1960s.

The sequence of nucleotides in the mRNA molecule is read in consecutive groups of three. RNA is a linear polymer of four different nucleotides, so there are $4 \times 4 \times 4 = 64$ possible combinations of three nucleotides: the triplets AAA, AUA, AUG, and so on. However, only 20 different amino acids are commonly found in proteins. Either some nucleotide triplets are never used, or the code is redundant and some amino acids are specified by more than one triplet. The second possibility is, in fact, the correct one, as shown by the completely deciphered genetic code in **Figure 6–50**. Each group of three consecutive nucleotides in RNA is called a **codon**, and each codon specifies either one amino acid or a stop to the translation process.

This genetic code is used universally in all present-day organisms. Although a few slight differences in the code have been found, these are chiefly in the DNA of mitochondria. Mitochondria have their own transcription and protein synthesis systems that operate quite independently from those of the rest of the cell, and it is understandable that their small genomes have been able to accommodate minor changes to the code (discussed in Chapter 14).

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AGA | | | | | | | | UUA | | | | | | AGC | | | | |
| | AGG | | | | | | | | UUG | | | | | | AGU | | | | |
| GCA | CGA | | | | | | GGA | | CUA | | | | | CCA | UCA | ACA | | | GUA | |
| GCC | CGC | | | | | | GGC | AUA | CUC | | | | | CCC | UCC | ACC | | | GUC | UAA |
| GCG | CGG | GAC | AAC | UGC | GAA | CAA | GGG | CAC | AUC | CUG | AAA | | UUC | CCG | UCG | ACG | | UAC | GUG | UAG |
| GCU | CGU | GAU | AAU | UGU | GAG | CAG | GGU | CAU | AUU | CUU | AAG | AUG | UUU | CCU | UCU | ACU | UGG | UAU | GUU | UGA |
| Ala | Arg | Asp | Asn | Cys | Glu | Gln | Gly | His | Ile | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val | stop |
| A | R | D | N | C | E | Q | G | H | I | L | K | M | F | P | S | T | W | Y | V | |

**Figure 6–50 The genetic code.** The standard one-letter abbreviation for each amino acid is presented below its three-letter abbreviation (see Panel 3–1, pp. 128–129, for the full name of each amino acid and its structure). By convention, codons are always written with the 5′-terminal nucleotide to the left. Note that most amino acids are represented by more than one codon, and that there are some regularities in the set of codons that specifies each amino acid. Codons for the same amino acid tend to contain the same nucleotides at the first and second positions, and vary at the third position. Three codons do not specify any amino acid but act as termination sites (stop codons), signaling the end of the protein-coding sequence. One codon—AUG—acts both as an initiation codon, signaling the start of a protein-coding message, and also as the codon that specifies methionine.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 114 of 240

In principle, an RNA sequence can be translated in any one of three different **reading frames**, depending on where the decoding process begins (**Figure 6–51**). However, only one of the three possible reading frames in an mRNA encodes the required protein. We see later how a special punctuation signal at the beginning of each RNA message sets the correct reading frame at the start of protein synthesis.

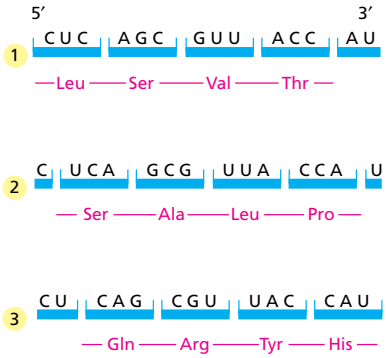



**Figure 6–51 The three possible reading frames in protein synthesis.** In the process of translating a nucleotide sequence *(blue)* into an amino acid sequence *(red),* the sequence of nucleotides in an mRNA molecule is read from the 5′ end to the 3′ end in consecutive sets of three nucleotides. In principle, therefore, the same RNA sequence can specify three completely different amino acid sequences, depending on the reading frame. In reality, however, only one of these reading frames contains the actual message.

## tRNA Molecules Match Amino Acids to Codons in mRNA

The codons in an mRNA molecule do not directly recognize the amino acids they specify: the group of three nucleotides does not, for example, bind directly to the amino acid. Rather, the translation of mRNA into protein depends on adaptor molecules that can recognize and bind both to the codon and, at another site on their surface, to the amino acid. These adaptors consist of a set of small RNA molecules known as **transfer RNAs** (**tRNAs**), each about 80 nucleotides in length.

We saw earlier in this chapter that RNA molecules can fold into precise three-dimensional structures, and the tRNA molecules provide a striking example. Four short segments of the folded tRNA are double-helical, producing a molecule that looks like a cloverleaf when drawn schematically (**Figure 6–52**). For example, a 5′-GCUC-3′ sequence in one part of a polynucleotide chain can form a relatively strong association with a 5′-GAGC-3′ sequence in another region of the same molecule. The cloverleaf undergoes further folding to form a compact L-shaped structure that is held together by additional hydrogen bonds between different regions of the molecule.

Two regions of unpaired nucleotides situated at either end of the L-shaped molecule are crucial to the function of tRNA in protein synthesis. One of these

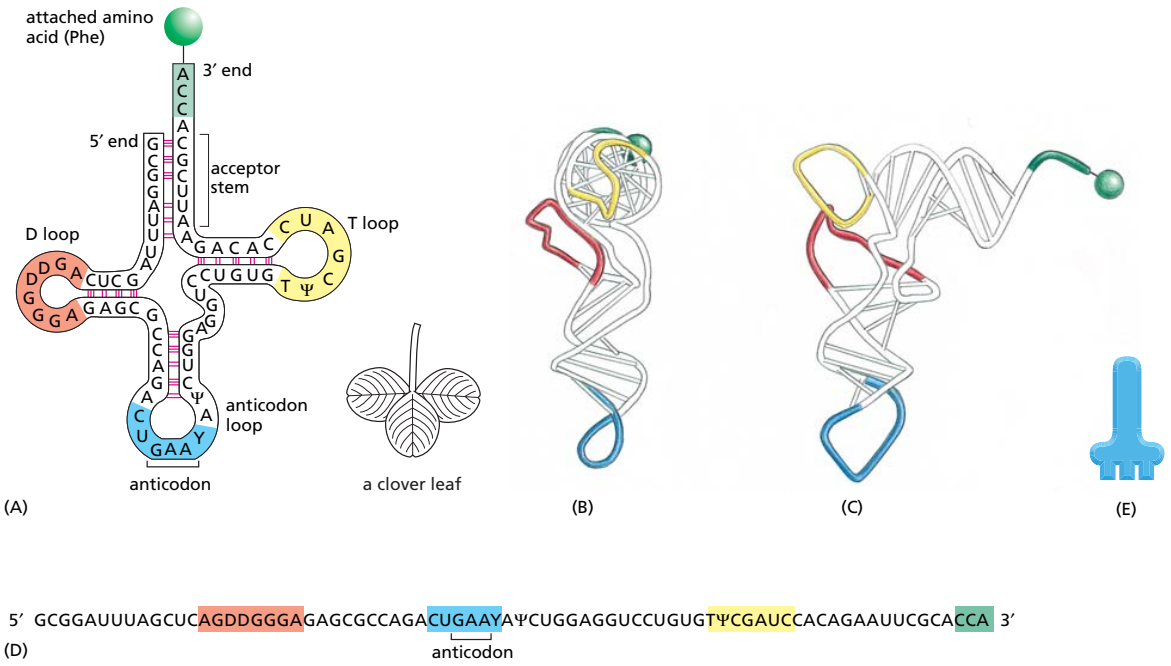



**Figure 6–52 A tRNA molecule.** A tRNA specific for the amino acid phenylalanine (Phe) is depicted in various ways. (A) The cloverleaf structure showing the complementary base-pairing *(red lines)* that creates the double-helical regions of the molecule. The anticodon is the sequence of three nucleotides that base-pairs with a codon in mRNA. The amino acid matching the codon/anticodon pair is attached at the 3′ end of the tRNA. tRNAs contain some unusual bases, which are produced by chemical modification after the tRNA has been synthesized. For example, the bases denoted ψ (pseudouridine—see Figure 6–43) and D (dihydrouridine—see Figure 6–55) are derived from uracil. (B and C) Views of the L-shaped molecule, based on x-ray diffraction analysis. Although this diagram shows the tRNA for the amino acid phenylalanine, all other tRNAs have similar structures. <CGCA> (D) The linear nucleotide sequence of the molecule, color-coded to match (A), (B), and (C). (E) The tRNA icon we use in this book.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 115 of 240

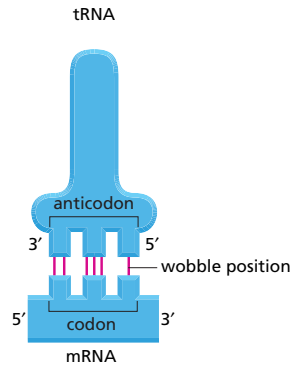
regions forms the **anticodon**, a set of three consecutive nucleotides that pairs with the complementary codon in an mRNA molecule. The other is a short single-stranded region at the 3′ end of the molecule; this is the site where the amino acid that matches the codon is attached to the tRNA.

We have seen in the previous section that the genetic code is redundant; that is, several different codons can specify a single amino acid (see Figure 6–50). This redundancy implies either that there is more than one tRNA for many of the amino acids or that some tRNA molecules can base-pair with more than one codon. In fact, both situations occur. Some amino acids have more than one tRNA and some tRNAs are constructed so that they require accurate base-pairing only at the first two positions of the codon and can tolerate a mismatch (or *wobble*) at the third position (**Figure 6–53**). This wobble base-pairing explains why so many of the alternative codons for an amino acid differ only in their third nucleotide (see Figure 6–50). In bacteria, wobble base-pairings make it possible to fit the 20 amino acids to their 61 codons with as few as 31 kinds of tRNA molecules. The exact number of different kinds of tRNAs, however, differs from one species to the next. For example, humans have nearly 500 tRNA genes but, among them, only 48 different anticodons are represented.

## tRNAs Are Covalently Modified Before They Exit from the Nucleus

Like most other eucaryotic RNAs, tRNAs are covalently modified before they are allowed to exit from the nucleus. Eucaryotic tRNAs are synthesized by RNA polymerase III. Both bacterial and eucaryotic tRNAs are typically synthesized as larger precursor tRNAs, which are then trimmed to produce the mature tRNA. In addition, some tRNA precursors (from both bacteria and eucaryotes) contain introns that must be spliced out. This splicing reaction differs chemically from pre-mRNA splicing; rather than generating a lariat intermediate, tRNA splicing uses a cut-and-paste mechanism that is catalyzed by proteins (**Figure 6–54**). Trimming and splicing both require the precursor tRNA to be correctly folded in its cloverleaf configuration. Because misfolded tRNA precursors will not be processed properly, the trimming and splicing reactions are thought to act as quality-control steps in the generation of tRNAs.

All tRNAs are modified chemically—nearly 1 in 10 nucleotides in each mature tRNA molecule is an altered version of a standard G, U, C, or A ribonucleotide. Over 50 different types of tRNA modifications are known; a few are shown in **Figure 6–55**. Some of the modified nucleotides—most notably inosine, produced by the deamination of adenosine—affect the conformation and base-pairing of the anticodon and thereby facilitate the recognition of the appropriate mRNA codon by the tRNA molecule (see Figure 6–53). Others affect the accuracy with which the tRNA is attached to the correct amino acid.




bacteria

| wobble codon base | possible anticodon bases |
|---|---|
| U | A, G, or I |
| C | G or I |
| A | U or I |
| G | C or U |

eucaryotes

| wobble codon base | possible anticodon bases |
|---|---|
| U | A, G, or I |
| C | G or I |
| A | U |
| G | C |

**Figure 6–53 Wobble base-pairing between codons and anticodons.** If the nucleotide listed in the first column is present at the third, or wobble, position of the codon, it can base-pair with any of the nucleotides listed in the second column. Thus, for example, when inosine (I) is present in the wobble position of the tRNA anticodon, the tRNA can recognize any one of three different codons in bacteria and either of two codons in eucaryotes. The inosine in tRNAs is formed from the deamination of guanine (see Figure 6–55), a chemical modification that takes place after the tRNA has been synthesized. The nonstandard base pairs, including those made with inosine, are generally weaker than conventional base pairs. Note that codon–anticodon base pairing is more stringent at positions 1 and 2 of the codon: here only conventional base pairs are permitted. The differences in wobble base-pairing interactions between bacteria and eucaryotes presumably result from subtle structural differences between bacterial and eucaryotic ribosomes, the molecular machines that perform protein synthesis. (Adapted from C. Guthrie and J. Abelson, in The Molecular Biology of the Yeast *Saccharomyces*: Metabolism and Gene Expression, pp. 487–528. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press, 1982.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 116 of 240



**Figure 6–54 Structure of a tRNA-splicing endonuclease docked to a precursor tRNA.** The endonuclease (a four-subunit enzyme) removes the tRNA intron *(blue)*. A second enzyme, a multifunctional tRNA ligase (not shown), then joins the two tRNA halves together. (Courtesy of Hong Li, Christopher Trotta and John Abelson.)

## Specific Enzymes Couple Each Amino Acid to Its Appropriate tRNA Molecule

We have seen that, to read the genetic code in DNA, cells make a series of different tRNAs. We now consider how each tRNA molecule becomes linked to the one amino acid in 20 that is its appropriate partner. Recognition and attachment of the correct amino acid depends on enzymes called **aminoacyl-tRNA synthetases**, which covalently couple each amino acid to its appropriate set of tRNA molecules (**Figure 6–56** and **Figure 6–57**). Most cells have a different synthetase enzyme for each amino acid (that is, 20 synthetases in all); one attaches glycine to all tRNAs that recognize codons for glycine, another attaches alanine to all tRNAs that recognize codons for alanine, and so on. Many bacteria, however, have fewer than 20 synthetases, and the same synthetase enzyme is responsible for coupling more than one amino acid to the appropriate tRNAs. In these cases, a single synthetase places the identical amino acid on two different types of tRNAs, only one of which has an anticodon that matches the amino acid. A second enzyme then chemically modifies each "incorrectly" attached amino acid so that it now corresponds to the anticodon displayed by its covalently linked tRNA.

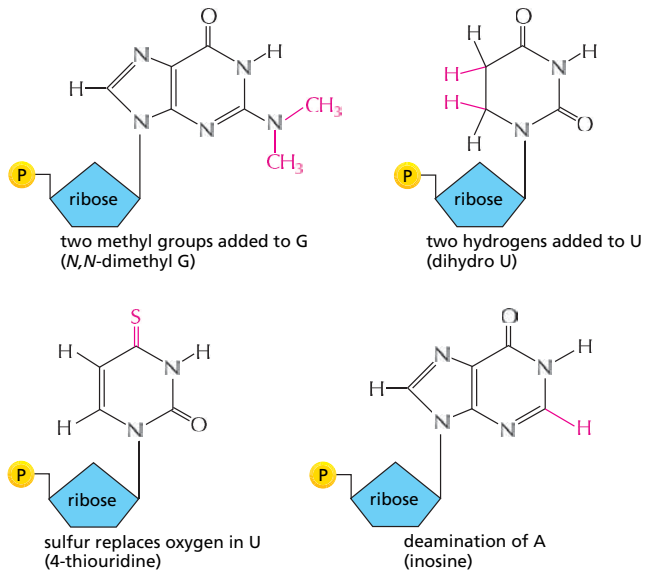



**Figure 6–55 A few of the unusual nucleotides found in tRNA molecules.** These nucleotides are produced by covalent modification of a normal nucleotide after it has been incorporated into an RNA chain. Two other types of modified nucleotides are shown in Figure 6–43. In most tRNA molecules about 10% of the nucleotides are modified (see Figure 6–52).

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 117 of 240

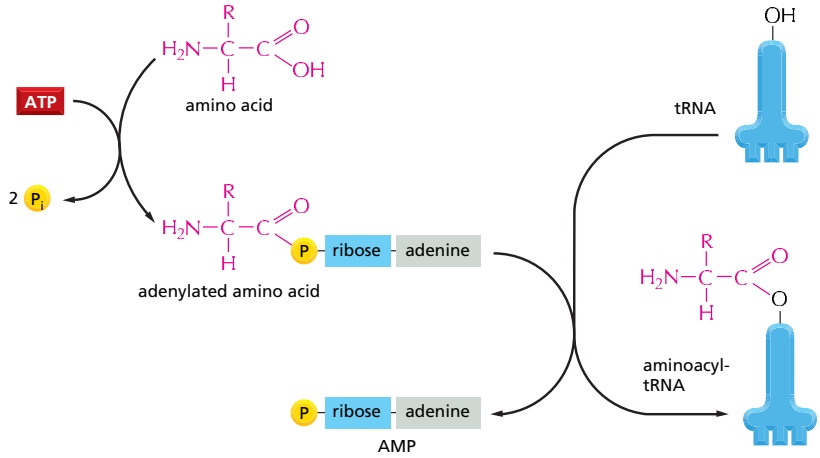



**Figure 6–56 Amino acid activation.** An amino acid is activated for protein synthesis by an aminoacyl-tRNA synthetase enzyme in two steps. As indicated, the energy of ATP hydrolysis is used to attach each amino acid to its tRNA molecule in a high-energy linkage. The amino acid is first activated through the linkage of its carboxyl group directly to an AMP moiety, forming an *adenylated amino acid*; the linkage of the AMP, normally an unfavorable reaction, is driven by the hydrolysis of the ATP molecule that donates the AMP. Without leaving the synthetase enzyme, the AMP-linked carboxyl group on the amino acid is then transferred to a hydroxyl group on the sugar at the 3′ end of the tRNA molecule. This transfer joins the amino acid by an activated ester linkage to the tRNA and forms the final aminoacyl-tRNA molecule. The synthetase enzyme is not shown in this diagram.

The synthetase-catalyzed reaction that attaches the amino acid to the 3′ end of the tRNA is one of many reactions coupled to the energy-releasing hydrolysis of ATP (see pp. 79–81), and it produces a high-energy bond between the tRNA and the amino acid. The energy of this bond is used at a later stage in protein synthesis to link the amino acid covalently to the growing polypeptide chain.

The aminoacyl-tRNA synthetase enzymes and the tRNAs are equally important in the decoding process (**Figure 6–58**). This was established by an experiment in which one amino acid (cysteine) was chemically converted into a different amino acid (alanine) after it already had been attached to its specific tRNA. When such "hybrid" aminoacyl-tRNA molecules were used for protein synthesis in a cell-free system, the wrong amino acid was inserted at every point in the protein chain where that tRNA was used. Although, as we shall see, cells have several quality control mechanisms to avoid this type of mishap, the experiment establishes that the genetic code is translated by two sets of adaptors that act sequentially. Each matches one molecular surface to another with great specificity, and it is their combined action that associates each sequence of three nucleotides in the mRNA molecule—that is, each codon—with its particular amino acid.

## Editing by tRNA Synthetases Ensures Accuracy

Several mechanisms working together ensure that the tRNA synthetase links the correct amino acid to each tRNA. The synthetase must first select the correct amino acid, and most synthetases do so by a two-step mechanism. First, the correct amino acid has the highest affinity for the active-site pocket of its synthetase

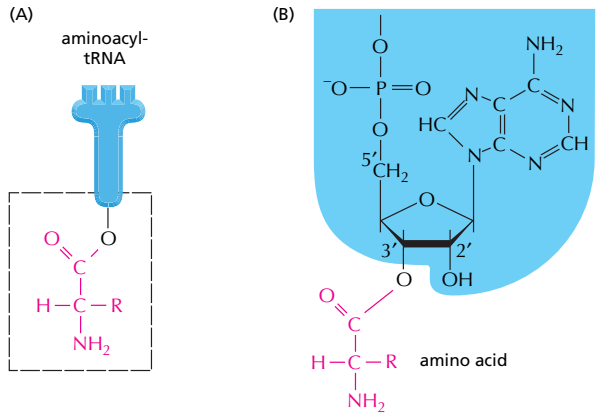
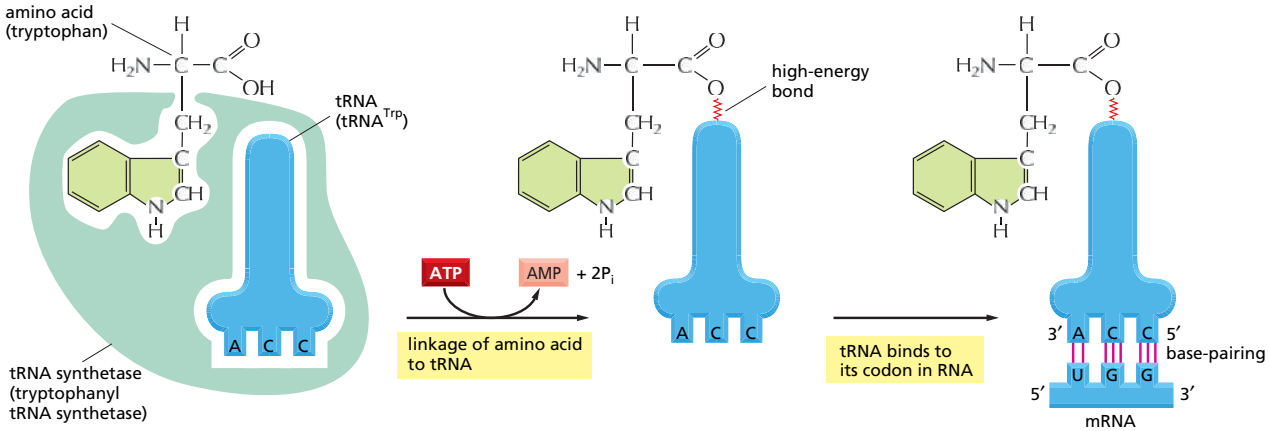



**Figure 6–57 The structure of the aminoacyl-tRNA linkage.** The carboxyl end of the amino acid forms an ester bond to ribose. Because the hydrolysis of this ester bond is associated with a large favorable change in free energy, an amino acid held in this way is said to be activated. (A) Schematic drawing of the structure. The amino acid is linked to the nucleotide at the 3′ end of the tRNA (see Figure 6–52). (B) Actual structure corresponding to the boxed region in (A). There are two major classes of synthetase enzymes: one links the amino acid directly to the 3′-OH group of the ribose, and the other links it initially to the 2′-OH group. In the latter case, a subsequent transesterification reaction shifts the amino acid to the 3′ position. As in Figure 6–56, the "R group" indicates the side chain of the amino acid.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 118 of 240




and is therefore favored over the other 19. In particular, amino acids larger than the correct one are effectively excluded from the active site. However, accurate discrimination between two similar amino acids, such as isoleucine and valine (which differ by only a methyl group), is very difficult to achieve by a one-step recognition mechanism. A second discrimination step occurs after the amino acid has been covalently linked to AMP (see Figure 6–56). When tRNA binds the synthetase, it tries to force the amino acid into a second pocket in the synthetase, the precise dimensions of which exclude the correct amino acid but allow access by closely related amino acids. Once an amino acid enters this editing pocket, it is hydrolyzed from the AMP (or from the tRNA itself if the aminoacyl-tRNA bond has already formed), and is released from the enzyme. This hydrolytic editing, which is analogous to the exonucleolytic proofreading by DNA polymerases (**Figure 6–59**), raises the overall accuracy of tRNA charging to approximately one mistake in 40,000 couplings.

**Figure 6–58 The genetic code is translated by means of two adaptors that act one after another.** The first adaptor is the aminoacyl-tRNA synthetase, which couples a particular amino acid to its corresponding tRNA; the second adaptor is the tRNA molecule itself, whose *anticodon* forms base pairs with the appropriate *codon* on the mRNA. An error in either step would cause the wrong amino acid to be incorporated into a protein chain. In the sequence of events shown, the amino acid tryptophan (Trp) is selected by the codon UGG on the mRNA.

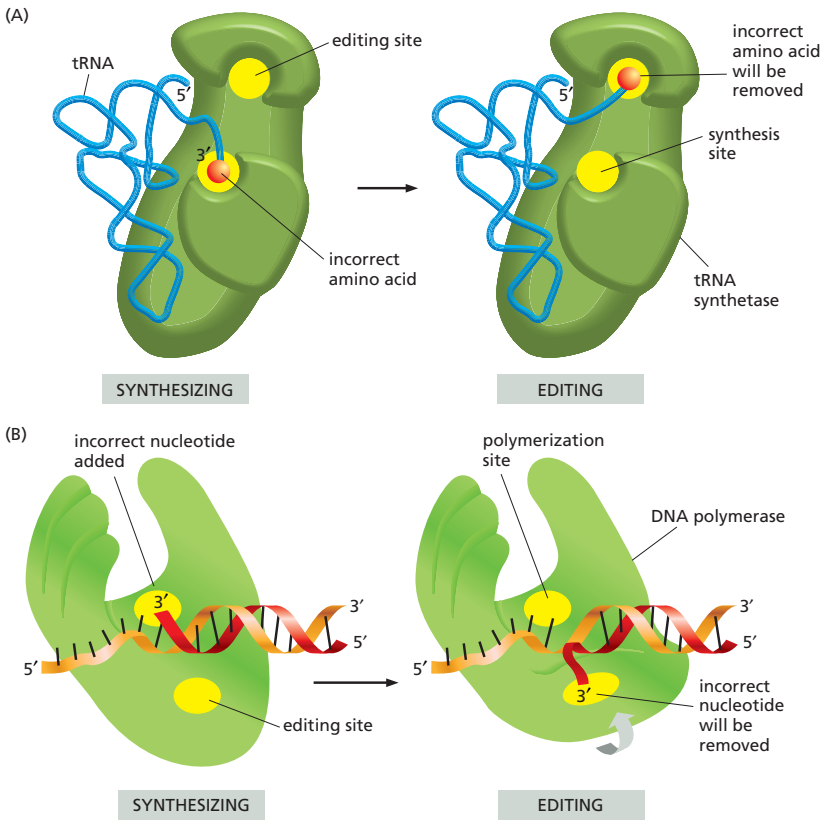



**Figure 6–59 Hydrolytic editing.** (A) tRNA synthetases remove their own coupling errors through hydrolytic editing of incorrectly attached amino acids. As described in the text, the correct amino acid is rejected by the editing site. (B) The error-correction process performed by DNA polymerase shows some similarities; however, it differs in so far as the removal process depends strongly on a mispairing with the template (see Figure 5–8).

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 119 of 240

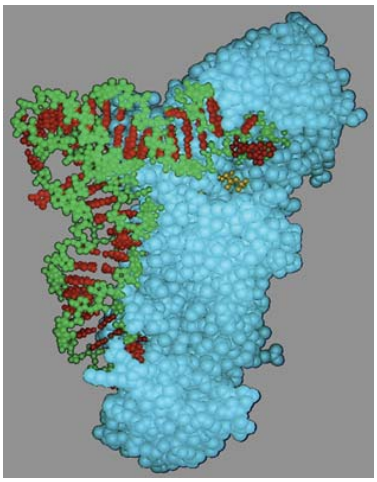


**Figure 6–60 The recognition of a tRNA molecule by its aminoacyl-tRNA synthetase.** For this tRNA ($tRNA^{Gln}$), specific nucleotides in both the anticodon *(bottom)* and the amino acid-accepting arm allow the correct tRNA to be recognized by the synthetase enzyme *(blue)*. A bound ATP molecule is *yellow*. (Courtesy of Tom Steitz.)

The tRNA synthetase must also recognize the correct set of tRNAs, and extensive structural and chemical complementarity between the synthetase and the tRNA allows the synthetase to probe various features of the tRNA (**Figure 6–60**). Most tRNA synthetases directly recognize the matching tRNA anticodon; these synthetases contain three adjacent nucleotide-binding pockets, each of which is complementary in shape and charge to a nucleotide in the anticodon. For other synthetases, the nucleotide sequence of the acceptor stem is the key recognition determinant. In most cases, however, the synthetase "reads" the nucleotides at several different positions on the tRNA.

## Amino Acids Are Added to the C-terminal End of a Growing Polypeptide Chain

Having seen that amino acids are first coupled to tRNA molecules, we now turn to the mechanism that joins amino acids together to form proteins. The fundamental reaction of protein synthesis is the formation of a peptide bond between the carboxyl group at the end of a growing polypeptide chain and a free amino group on an incoming amino acid. Consequently, a protein is synthesized stepwise from its N-terminal end to its C-terminal end. Throughout the entire process the growing carboxyl end of the polypeptide chain remains activated by its covalent attachment to a tRNA molecule (forming a peptidyl-tRNA). Each addition disrupts this high-energy covalent linkage, but immediately replaces it with an identical linkage on the most recently added amino acid (**Figure 6–61**). In this way, each amino acid added carries with it the activation energy for the addition of the next amino acid rather than the energy for its own addition—an example of the "head growth" type of polymerization described in Figure 2–68.

**Figure 6–61 The incorporation of an amino acid into a protein.** A polypeptide chain grows by the stepwise addition of amino acids to its C-terminal end. The formation of each peptide bond is energetically favorable because the growing C-terminus has been activated by the covalent attachment of a tRNA molecule. The peptidyl-tRNA linkage that activates the growing end is regenerated during each addition. The amino acid side chains have been abbreviated as $R_1$, $R_2$, $R_3$, and $R_4$; as a reference point, all of the atoms in the second amino acid in the polypeptide chain are shaded *gray*. The figure shows the addition of the fourth amino acid *(red)* to the growing chain.

## The RNA Message Is Decoded in Ribosomes

The synthesis of proteins is guided by information carried by mRNA molecules. To maintain the correct reading frame and to ensure accuracy (about 1 mistake every 10,000 amino acids), protein synthesis is performed in the **ribosome**, a complex catalytic machine made from more than 50 different proteins (the *ribosomal proteins*) and several RNA molecules, the **ribosomal RNAs** (**rRNAs**). <CGCC> A typical eucaryotic cell contains millions of ribosomes in its cytoplasm (**Figure 6–62**). Eucaryotic ribosome subunits are assembled at the nucleolus, when newly transcribed and modified rRNAs associate with ribosomal

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 120 of 240

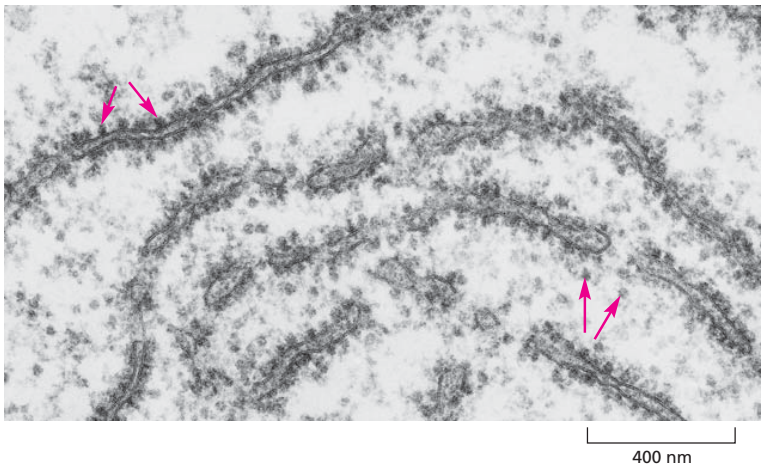



**Figure 6–62 Ribosomes in the cytoplasm of a eucaryotic cell.** This electron micrograph shows a thin section of a small region of cytoplasm. The ribosomes appear as black dots *(red arrows)*. Some are free in the cytosol; others are attached to membranes of the endoplasmic reticulum. (Courtesy of Daniel S. Friend.)

proteins, which have been transported into the nucleus after their synthesis in the cytoplasm. The two ribosomal subunits are then exported to the cytoplasm, where they join together to synthesize proteins.

Eucaryotic and procaryotic ribosomes have similar designs and functions. Both are composed of one large and one small subunit that fit together to form a complete ribosome with a mass of several million daltons (**Figure 6–63**). The small subunit provides the framework on which the tRNAs can be accurately

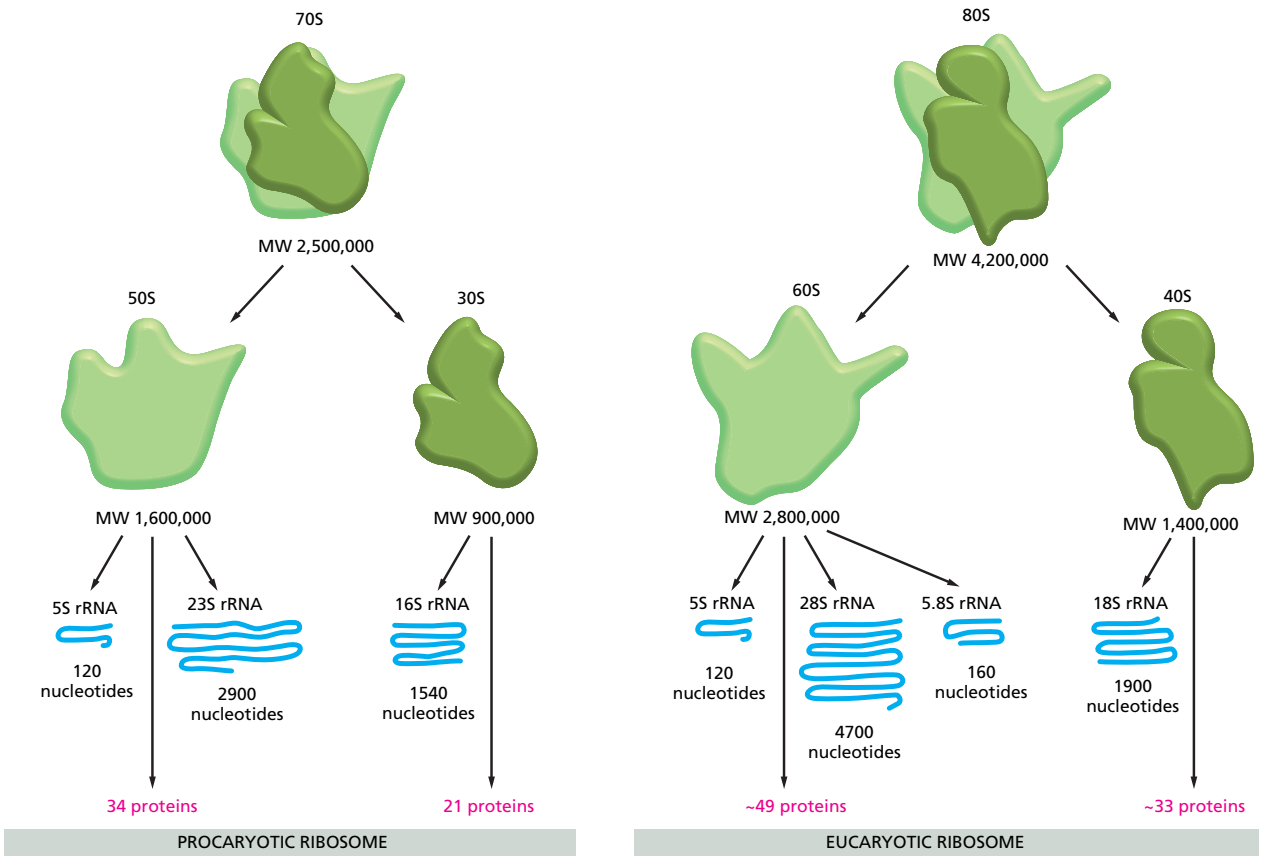



**Figure 6–63 A comparison of procaryotic and eucaryotic ribosomes.** Despite differences in the number and size of their rRNA and protein components, both procaryotic and eucaryotic ribosomes have nearly the same structure and they function similarly. Although the 18S and 28S rRNAs of the eucaryotic ribosome contain many nucleotides not present in their bacterial counterparts, these nucleotides are present as multiple insertions that form extra domains and leave the basic structure of each rRNA largely unchanged.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 121 of 240

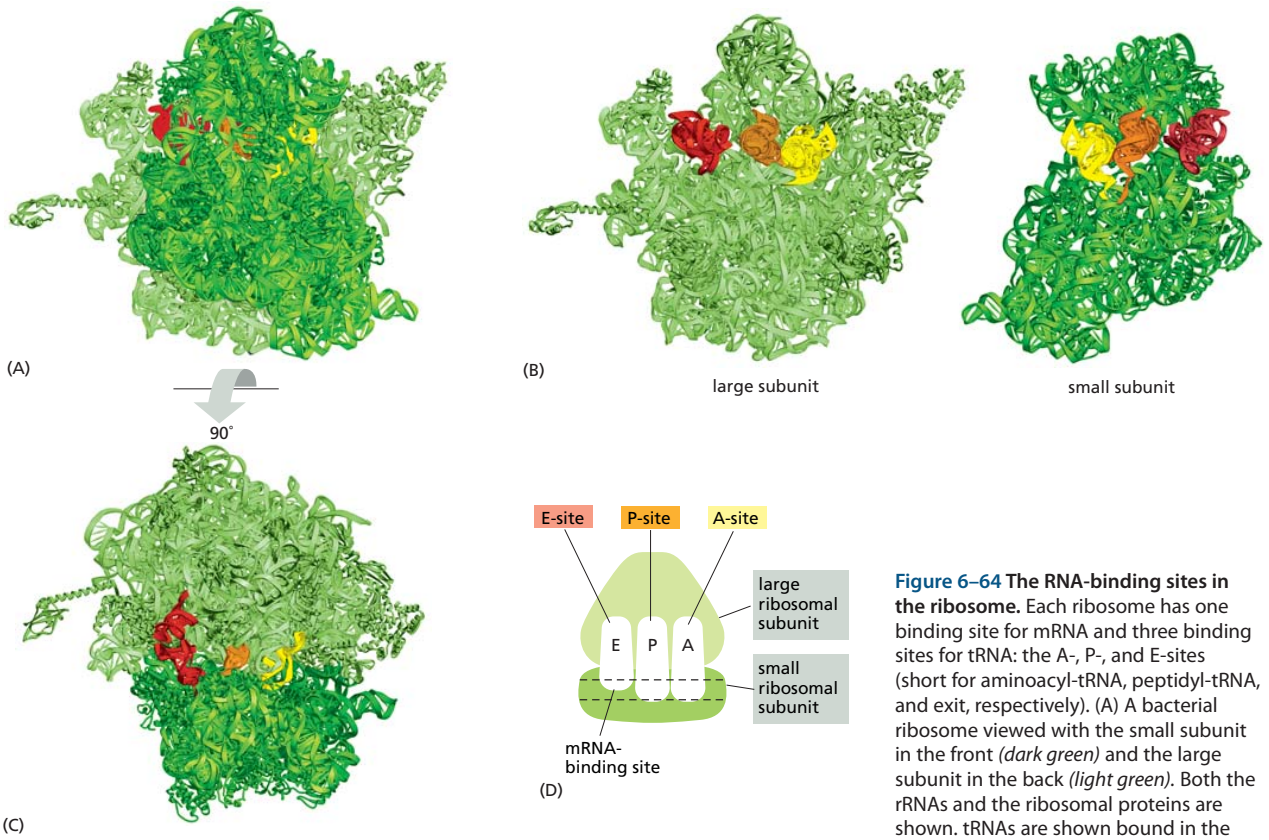



**Figure 6–64 The RNA-binding sites in the ribosome.** Each ribosome has one binding site for mRNA and three binding sites for tRNA: the A-, P-, and E-sites (short for aminoacyl-tRNA, peptidyl-tRNA, and exit, respectively). (A) A bacterial ribosome viewed with the small subunit in the front *(dark green)* and the large subunit in the back *(light green)*. Both the rRNAs and the ribosomal proteins are shown. tRNAs are shown bound in the E-site *(red)*, the P-site *(orange)* and the A-site *(yellow)*. Although all three tRNA sites are shown occupied here, during the process of protein synthesis not more than two of these sites are thought to contain tRNA molecules at any one time (see Figure 6–66). (B) Large and small ribosomal subunits arranged as though the ribosome in (A) were opened like a book. (C) The ribosome in (A) rotated through 90º and viewed with the large subunit on top and small subunit on the bottom. (D) Schematic representation of a ribosome (in the same orientation as C), which will be used in subsequent figures. (A, B, and C, adapted from M.M. Yusupov et al., *Science* 292:883–896, 2001. With permission from AAAS; courtesy of Albion Baucom and Harry Noller.)

matched to the codons of the mRNA (see Figure 6–58), while the large subunit catalyzes the formation of the peptide bonds that link the amino acids together into a polypeptide chain (see Figure 6–61).

When not actively synthesizing proteins, the two subunits of the ribosome are separate. They join together on an mRNA molecule, usually near its 5′ end, to initiate the synthesis of a protein. The mRNA is then pulled through the ribosome; as its codons enter the core of the ribosome, the mRNA nucleotide sequence is translated into an amino acid sequence using the tRNAs as adaptors to add each amino acid in the correct sequence to the end of the growing polypeptide chain. When a stop codon is encountered, the ribosome releases the finished protein, and its two subunits separate again. These subunits can then be used to start the synthesis of another protein on another mRNA molecule.

Ribosomes operate with remarkable efficiency: in one second, a single ribosome of a eucaryotic cell adds about 2 amino acids to a polypeptide chain; the ribosomes of bacterial cells operate even faster, at a rate of about 20 amino acids per second. How does the ribosome choreograph the many coordinated movements required for efficient translation? A ribosome contains four binding sites for RNA molecules: one is for the mRNA and three (called the A-site, the P-site, and the E-site) are for tRNAs (**Figure 6–64**). A tRNA molecule is held tightly at the A- and P-sites only if its anticodon forms base pairs with a complementary codon (allowing for wobble) on the mRNA molecule that is threaded through the ribosome (**Figure 6–65**). The A- and P-sites are close enough together for their two tRNA molecules to be forced to form base pairs with adjacent codons on the mRNA molecule. This feature of the ribosome maintains the correct reading frame on the mRNA.

Once protein synthesis has been initiated, each new amino acid is added to the elongating chain in a cycle of reactions containing four major steps: tRNA

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 122 of 240

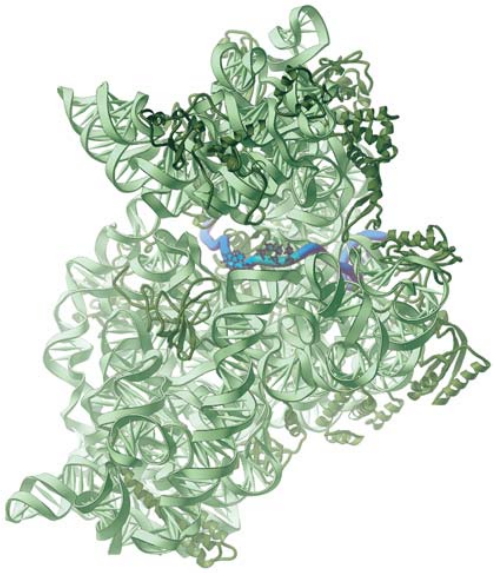


**Figure 6–65 The path of mRNA *(blue)* through the small ribosomal subunit.** The orientation is the same as that in the right-hand panel of Figure 6–64B. (Courtesy of Harry F. Noller, based on data in G.Z. Yusopova et al., *Cell* 106:233–241, 2001. With permission from Elsevier.)

binding, peptide bond formation, large subunit and small subunit translocation. As a result of the two translocation steps, the entire ribosome moves three nucleotides along the mRNA and is positioned to start the next cycle. (**Figure 6–66**). Our description of the chain elongation process begins at a point at which some amino acids have already been linked together and there is a tRNA molecule in the P-site on the ribosome, covalently joined to the end of the growing polypeptide. In step 1, a tRNA carrying the next amino acid in the chain binds to the ribosomal A-site by forming base pairs with the mRNA codon positioned there, so that the P-site and the A-site contain adjacent bound tRNAs. In step 2, the carboxyl end of the polypeptide chain is released from the tRNA at the P-site (by breakage of the high-energy bond between the tRNA and its amino acid) and joined to the free amino group of the amino acid linked to the tRNA at the A-site, forming a new peptide bond. This central reaction of protein synthesis is catalyzed by a *peptidyl transferase* contained in the large ribosomal subunit. In step 3, the large subunit moves relative to the mRNA held by the small subunit, thereby shifting the acceptor stems of the two tRNAs to the E- and P-sites of the large subunit. In step 4, another series of conformational changes moves the small subunit and its bound mRNA exactly three nucleotides, resetting the ribosome so it is ready to receive the next aminoacyl-tRNA. Step 1 is then repeated with a new incoming aminoacyl-tRNA, and so on. <CGTT>

This four-step cycle is repeated each time an amino acid is added to the polypeptide chain, as the chain grows from its amino to its carboxyl end.

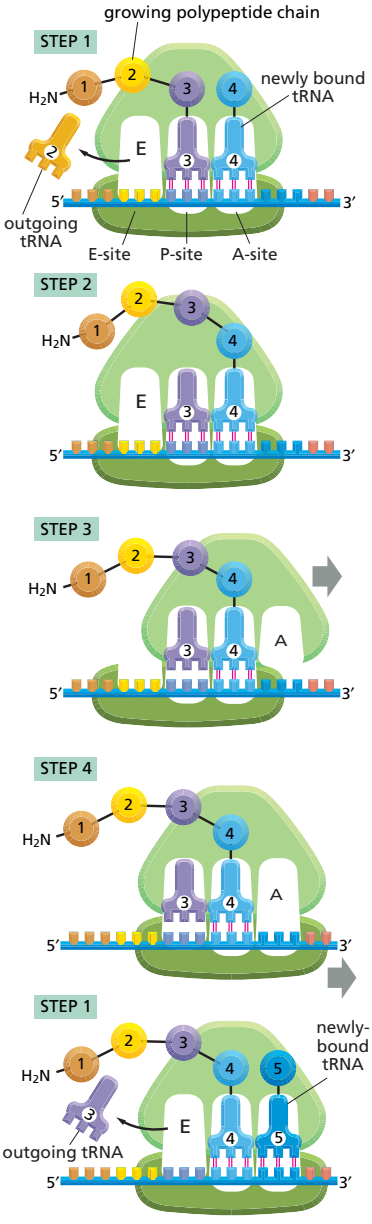



**Figure 6–66 Translating an mRNA molecule.** Each amino acid added to the growing end of a polypeptide chain is selected by complementary base-pairing between the anticodon on its attached tRNA molecule and the next codon on the mRNA chain. Because only one of the many types of tRNA molecules in a cell can base-pair with each codon, the codon determines the specific amino acid to be added to the growing polypeptide chain. The four-step cycle shown is repeated over and over during the synthesis of a protein. In step 1, an aminoacyl-tRNA molecule binds to a vacant A-site on the ribosome and a spent tRNA molecule dissociates from the E-site. In step 2, a new peptide bond is formed. In step 3, the large subunit translocates relative to the small subunit, leaving the two tRNAs in hybrid sites: P on the large subunit and A on the small, for one; E on the large subunit and P on the small, for the other. In step 4, the small subunit translocates carrying its mRNA a distance of three nucleotides through the ribosome. This "resets" the ribosome with a fully empty A-site, ready for the next aminoacyl-tRNA molecule to bind. As indicated, the mRNA is translated in the 5′-to-3′ direction, and the N-terminal end of a protein is made first, with each cycle adding one amino acid to the C-terminus of the polypeptide chain.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 123 of 240

## Elongation Factors Drive Translation Forward and Improve Its Accuracy

The basic cycle of polypeptide elongation shown in outline in Figure 6–66 has an additional feature that makes translation especially efficient and accurate. Two *elongation factors* enter and leave the ribosome during each cycle, each hydrolyzing GTP to GDP and undergoing conformational changes in the process. These factors are called EF-Tu and EF-G in bacteria, and EF1 and EF2 in eucaryotes. Under some conditions *in vitro,* ribosomes can be forced to synthesize proteins without the aid of these elongation factors and GTP hydrolysis, but this synthesis is very slow, inefficient, and inaccurate. Coupling the GTP hydrolysis-driven changes in the elongation factors to transitions between different states of the ribosome speeds up protein synthesis enormously. Although these ribosomal states are not yet understood in detail, they almost certainly involve RNA structure rearrangements in the ribosome core. The cycles of elongation factor association, GTP hydrolysis, and dissociation ensure that all such changes occur in the "forward" direction so that translation can proceed efficiently (**Figure 6–67**).

As shown previously, EF-Tu simultaneously binds GTP and aminoacyl-tRNAs (see Figure 3–74). In addition to helping move translation forward, EF-Tu (EF1 in eucaryotes) increases the accuracy of translation in several ways. First, as it escorts an incoming aminoacyl-tRNA to the ribosome, EF-Tu checks whether the tRNA–amino acid match is corrrect. Exactly how this is accomplished is not well understood. According to one idea, correct tRNA–amino acid matches have a narrowly defined affinity for EF-Tu, which allows EF-Tu to discriminate, albeit crudely, among many different amino acid–tRNA combinations, selectively bringing the correct ones with it into the ribosome. Second, EF-Tu monitors the initial interaction between the anticodon of an incoming aminoacyl-tRNA and the codon of the mRNA in the A-site. Aminoacyl-tRNAs are "bent" when bound to the GTP-form of EF-Tu; this bent conformation allows codon pairing but prevents incorporation of the amino acid into the growing polypeptide chain. However, if the codon–anticodon match is correct, the ribosome rapidly triggers the hydrolysis of the GTP molecule, whereupon EF-Tu releases its grip on the tRNA and dissociates from the ribosome, allowing the tRNA to donate its amino acid for protein synthesis. But how is the "correctness" of the codon–anticodon match assessed? This feat is carried out by the ribosome itself through an RNA-based mechanism. The rRNA in the small subunit of the ribosome forms a series of hydrogen bonds with the codon–anticodon pair that allows determination of its correctness (**Figure 6–68**). In essence, the rRNA folds around the codon–anticodon pair, and its final closure—which occurs only when the correct anticodon is in place—triggers GTP hydrolysis. Remarkably, this induced fit mechanism can distinguish correct from incorrect codon–anticodon interactions despite the rules for wobble base-pairing summarized in Figure 6–53. From this example, as for RNA splicing, one gets a sense of the highly sophisticated forms of molecular recognition that can be achieved solely by RNA.

The interactions of EF-Tu, tRNA, and the ribosome just described introduce critical proofreading steps into protein synthesis at the initial tRNA selection stage. But after GTP is hydrolyzed and EF-Tu dissociates from the ribosome, there is an additional opportunity for the ribosome to prevent an incorrect amino acid from being added to the growing chain. Following GTP hydrolysis, there is a short time delay as the amino acid carried by the tRNA moves into position on the ribosome. This time delay is shorter for correct than incorrect codon–anticodon pairs. Moreover, incorrectly matched tRNAs dissociate more

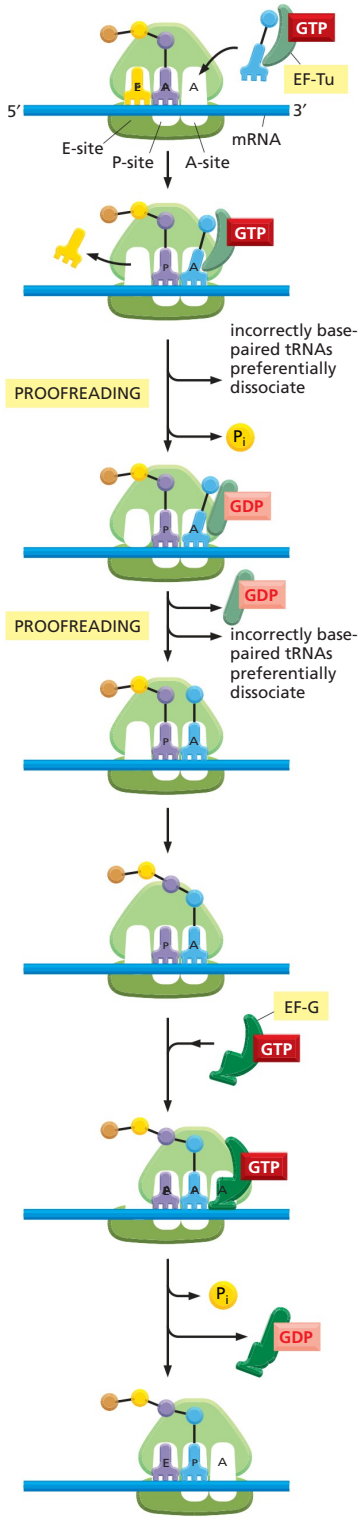



**Figure 6–67 Detailed view of the translation cycle.** The outline of translation presented in Figure 6–66 has been expanded to show the roles of two elongation factors EF-Tu and EF-G, which drive translation in the forward direction. As explained in the text, EF-Tu also provides two opportunities for proofreading of the codon–anticodon match. In this way, incorrectly paired tRNAs are selectively rejected, and the accuracy of translation is improved.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 124 of 240

rapidly than those correctly bound because their interaction with the codon is weaker. Thus, most incorrectly bound tRNA molecules (as well as a significant number of correctly bound molecules) will leave the ribosome without being used for protein synthesis. All of these proofreading steps, taken together, are largely responsible for the 99.99% accuracy of the ribosome in translating RNA into protein.

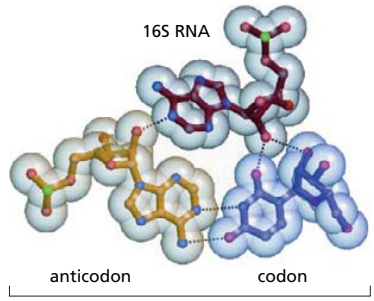



**Figure 6–68 Recognition of correct codon–anticodon matches by the small subunit rRNA of the ribosome.** Shown is the interaction between a nucleotide of the small subunit rRNA and the first nucleotide pair of a correctly paired codon–anticodon; similar interactions occur between other nucleotides of the rRNA and the second and third positions of the codon–anticodon pair. The small-subunit rRNA can form this network of hydrogen bonds only with correctly matched codon–anticodon pairs. As explained in the text, this codon–anticodon monitoring by the small-subunit rRNA increases the accuracy of protein synthesis. (From J.M. Ogle et al., *Science* 292:897–902, 2001. With permission from AAAS.)

## The Ribosome Is a Ribozyme

The ribosome is a large complex composed of two-thirds RNA and one-third protein. The determination, in 2000, of the entire three-dimensional conformation of its large and small subunits is a major triumph of modern structural biology. The findings confirm earlier evidence that rRNAs—and not proteins—are responsible for the ribosome's overall structure, its ability to position tRNAs on the mRNA, and its catalytic activity in forming covalent peptide bonds. The ribosomal RNAs are folded into highly compact, precise three-dimensional structures that form the compact core of the ribosome and determine its overall shape (**Figure 6–69**).

In marked contrast to the central positions of the rRNAs, the ribosomal proteins are generally located on the surface and fill in the gaps and crevices of the folded RNA (**Figure 6–70**). Some of these proteins send out extended regions of polypeptide chain that penetrate short distances into holes in the RNA core (**Figure 6–71**). The main role of the ribosomal proteins seems to be to stabilize the RNA core, while permitting the changes in rRNA conformation that are necessary for this RNA to catalyze efficient protein synthesis. The proteins probably also aid in the initial assembly of the rRNAs that make up the core of the ribosome.

Not only are the A-, P-, and E-binding sites for tRNAs formed primarily by ribosomal RNAs, but the catalytic site for peptide bond formation is also formed by RNA, as the nearest amino acid is located more than 1.8 nm away.

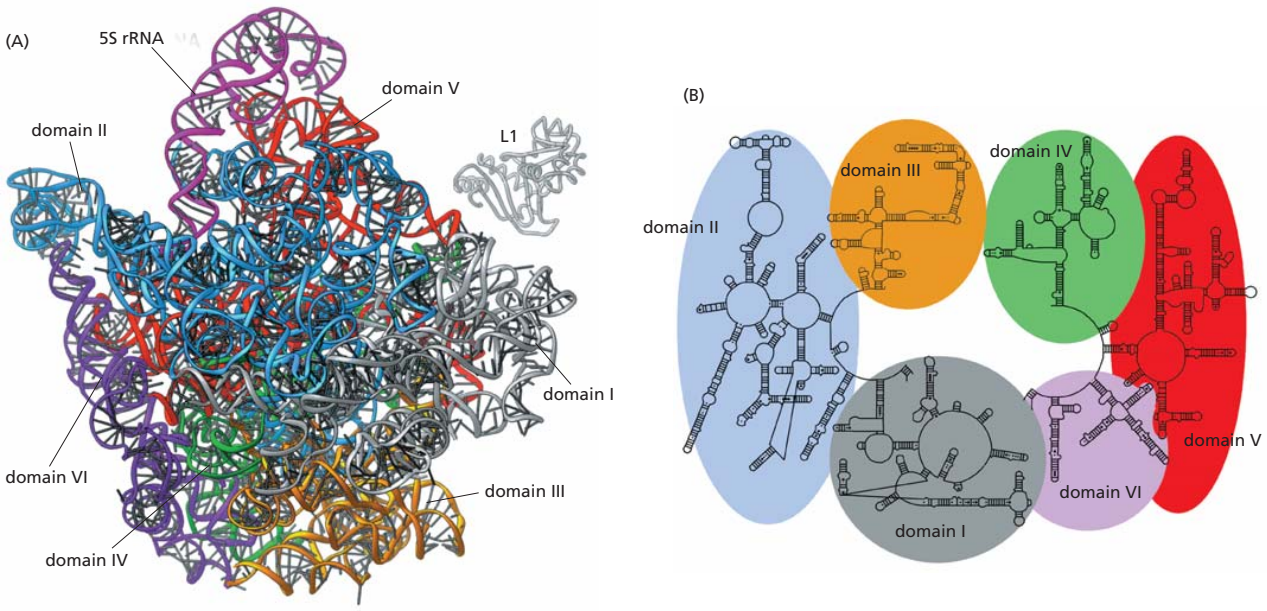



**Figure 6–69 Structure of the rRNAs in the large subunit of a bacterial ribosome, as determined by x-ray crystallography.** (A) Three-dimensional conformations of the large-subunit rRNAs (5S and 23S) as they appear in the ribosome. One of the protein subunits of the ribosome (L1) is also shown as a reference point, since it forms a characteristic protrusion on the ribosome. (B) Schematic diagram of the secondary structure of the 23S rRNA, showing the extensive network of base-pairing. The structure has been divided into six "domains" whose colors correspond to those in (A). The secondary-structure diagram is highly schematized to represent as much of the structure as possible in two dimensions. To do this, several discontinuities in the RNA chain have been introduced, although in reality the 23S RNA is a single RNA molecule. For example, the base of Domain III is continuous with the base of Domain IV even though a gap appears in the diagram. (Adapted from N. Ban et al., *Science* 289:905–920, 2000. With permission from AAAS.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 125 of 240

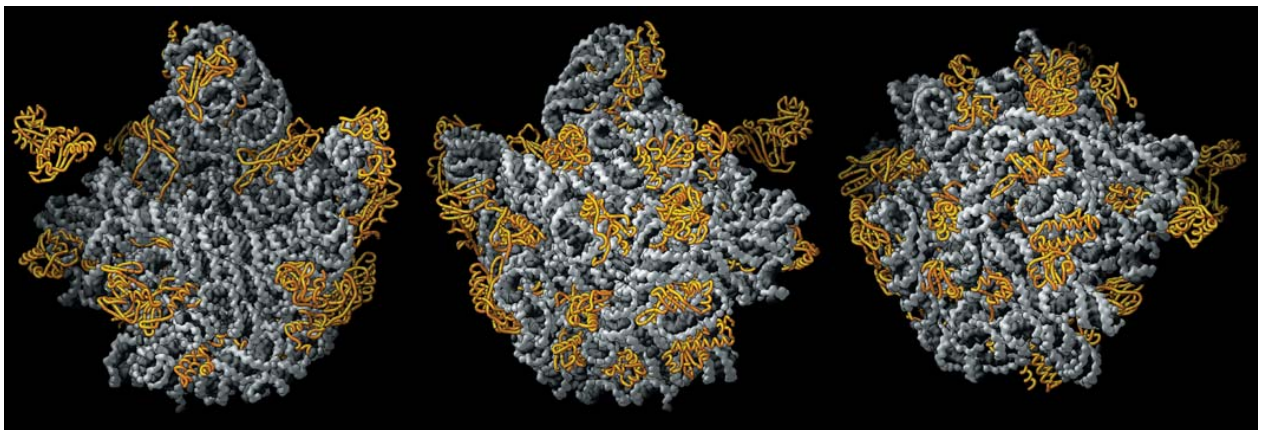



**Figure 6–70 Location of the protein components of the bacterial large ribosomal subunit.** The rRNAs (5S and 23S) are shown in *gray* and the large-subunit proteins (27 of the 31 total) in *gold*. For convenience, the protein structures depict only the polypeptide backbones. (A) Interface with the small subunit, the same view shown in Figure 6–64B. (B) Side opposite to that shown in (A), obtained by rotating (A) by 180° around a vertical axis. (C) Further slight rotation of (B) through a diagonal axis, allowing a view into the peptide exit channel in the center of the structure. (From N. Ban et al., *Science* 289:905–920, 2000. With permission from AAAS.)

This discovery came as a surprise to biologists because, unlike proteins, RNA does not contain easily ionizable functional groups that can be used to catalyze sophisticated reactions like peptide bond formation. Moreover, metal ions, which are often used by RNA molecules to catalyze chemical reactions (as discussed later in the chapter), were not observed at the active site of the ribosome. Instead, it is believed that the 23S rRNA forms a highly structured pocket that, through a network of hydrogen bonds, precisely orients the two reactants (the growing peptide chain and an aminoacyl-tRNA) and thereby greatly accelerates their covalent joining. In addition, the tRNA in the P site contributes to the active site, perhaps supplying a functional OH group that participates directly in the catalysis. This mechanism may ensure that catalysis occurs only when the tRNA is properly positioned in the ribosome.

RNA molecules that possess catalytic activity are known as **ribozymes**. We saw earlier in this chapter how other ribozymes function in self-splicing reactions (for example, see Figure 6–36). In the final section of this chapter, we consider what the ability of RNA molecules to function as catalysts for a wide variety of different reactions might mean for the early evolution of living cells. For now, we merely note that there is good reason to suspect that RNA rather than protein molecules served as the first catalysts for living cells. If so, the ribosome, with its RNA core, may be a relic of an earlier time in life's history—when protein synthesis evolved in cells that were run almost entirely by ribozymes.



**Figure 6–71 Structure of the L15 protein in the large subunit of the bacterial ribosome.** The globular domain of the protein lies on the surface of the ribosome and an extended region penetrates deeply into the RNA core of the ribosome. The L15 protein is shown in *yellow* and a portion of the ribosomal RNA core is shown in *red*. (From D. Klein, P.B. Moore and T.A. Steitz, *J. Mol. Biol.* 340:141–147, 2004. With permission from Academic Press.)

## Nucleotide Sequences in mRNA Signal Where to Start Protein Synthesis

The initiation and termination of translation share features of the translation elongation cycle described above. The site at which protein synthesis begins on the mRNA is especially crucial, since it sets the reading frame for the whole length of the message. An error of one nucleotide either way at this stage would cause every subsequent codon in the message to be misread, resulting in a nonfunctional protein with a garbled sequence of amino acids. The initiation step is also important because for most genes it is the last point at which the cell can decide whether the mRNA is to be translated and the protein synthesized; the rate of initiation is thus one determinant of the rate at which any protein is synthesized. We shall see in Chapter 7 that cells use several mechanisms to regulate translation initiation.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 126 of 240

The translation of an mRNA begins with the codon AUG, and a special tRNA is required to start translation. This **initiator tRNA** always carries the amino acid methionine (in bacteria, a modified form of methionine—formylmethionine—is used), with the result that all newly made proteins have methionine as the first amino acid at their N-terminus, the end of a protein that is synthesized first. This methionine is usually removed later by a specific protease. The initiator tRNA can be specially recognized by initiation factors because it has a nucleotide sequence distinct from that of the tRNA that normally carries methionine.

In eucaryotes, the initiator tRNA–methionine complex (Met–tRNAi) is first loaded into the small ribosomal subunit along with additional proteins called **eucaryotic initiation factors**, or **eIFs** (**Figure 6–72**). Of all the aminoacyl-tRNAs in the cell, only the methionine-charged initiator tRNA is capable of tightly binding the small ribosome subunit without the complete ribosome being present and it binds directly to the P-site. Next, the small ribosomal subunit binds to the 5′ end of an mRNA molecule, which is recognized by virtue of its 5′ cap and its two bound initiation factors, eIF4E (which directly binds the cap) and eIF4G (see Figure 6–40). The small ribosomal subunit then moves forward (5′ to 3′) along the mRNA, searching for the first AUG. Additional initiation factors that act as ATP-powered helicases facilitate the ribosome's movement through RNA secondary structure. In 90% of mRNAs, translation begins at the first AUG encountered by the small subunit. At this point, the initiation factors dissociate, allowing the large ribosomal subunit to assemble with the complex and complete the ribosome. The initiator tRNA is still bound to the P-site, leaving the A-site vacant. Protein synthesis is therefore ready to begin (see Figure 6–72).

The nucleotides immediately surrounding the start site in eucaryotic mRNAs influence the efficiency of AUG recognition during the above scanning process. If this recognition site differs substantially from the consensus recognition sequence (5′-ACCAUGG-3′), scanning ribosomal subunits will sometimes ignore the first AUG codon in the mRNA and skip to the second or third AUG codon instead. Cells frequently use this phenomenon, known as "leaky scanning," to produce two or more proteins, differing in their N-termini, from the same mRNA molecule. It allows some genes to produce the same protein with and without a signal sequence attached at its N-terminus, for example, so that the protein is directed to two different compartments in the cell.

The mechanism for selecting a start codon in bacteria is different. Bacterial mRNAs have no 5′ caps to signal the ribosome where to begin searching for the start of translation. Instead, each bacterial mRNA contains a specific ribosome-binding site (called the Shine–Dalgarno sequence, named after its discoverers) that is located a few nucleotides upstream of the AUG at which translation is to begin. This nucleotide sequence, with the consensus 5′-AGGAGGU-3′, forms base pairs with the 16S rRNA of the small ribosomal subunit to position the initiating AUG codon in the ribosome. A set of translation initiation factors orchestrates this interaction, as well as the subsequent assembly of the large ribosomal subunit to complete the ribosome.

Unlike a eucaryotic ribosome, a bacterial ribosome can therefore readily assemble directly on a start codon that lies in the interior of an mRNA molecule, so long as a ribosome-binding site precedes it by several nucleotides. As a result, bacterial mRNAs are often *polycistronic*—that is, they encode several different proteins, each of which is translated from the same mRNA molecule (**Figure 6–73**). In contrast, a eucaryotic mRNA generally encodes only a single protein.

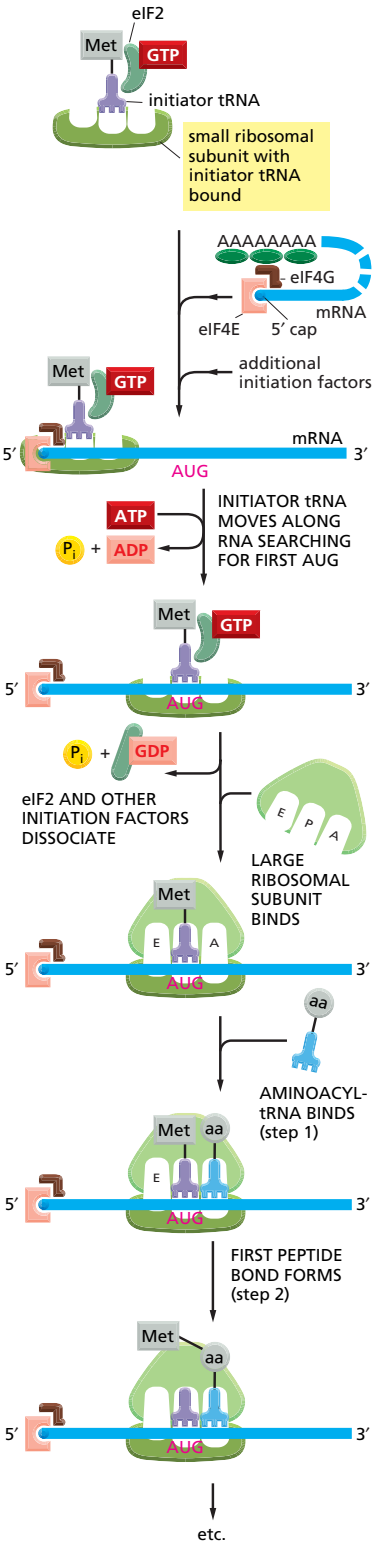



**Figure 6–72 The initiation of protein synthesis in eucaryotes.** Only three of the many translation initiation factors required for this process are shown. Efficient translation initiation also requires the poly-A tail of the mRNA bound by poly-A-binding proteins which, in turn, interact with eIF4G. In this way, the translation apparatus ascertains that both ends of the mRNA are intact before initiating protein synthesis (see Figure 6–40). Although only one GTP hydrolysis event is shown in the figure, a second is known to occur just before the large and small ribosomal subunits join.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 127 of 240

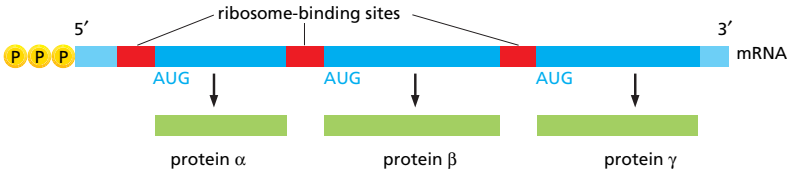



**Figure 6–73 Structure of a typical bacterial mRNA molecule.** Unlike eucaryotic ribosomes, which typically require a capped 5′ end, procaryotic ribosomes initiate transcription at ribosome-binding sites (Shine–Dalgarno sequences), which can be located anywhere along an mRNA molecule. This property of ribosomes permits bacteria to synthesize more than one type of protein from a single mRNA molecule.

## Stop Codons Mark the End of Translation

The end of the protein-coding message is signaled by the presence of one of three *stop codons* (UAA, UAG, or UGA) (see Figure 6–50). These are not recognized by a tRNA and do not specify an amino acid, but instead signal to the ribosome to stop translation. Proteins known as *release factors* bind to any ribosome with a stop codon positioned in the A site, forcing the peptidyl transferase in the ribosome to catalyze the addition of a water molecule instead of an amino acid to the peptidyl-tRNA (**Figure 6–74**). This reaction frees the carboxyl end of the growing polypeptide chain from its attachment to a tRNA molecule, and since only this attachment normally holds the growing polypeptide to the ribosome, the completed protein chain is immediately released into the cytoplasm. The ribosome then releases the mRNA and separates into the large and small subunits, which can assemble on this or another mRNA molecule to begin a new round of protein synthesis.

Release factors are an example of *molecular mimicry*, whereby one type of macromolecule resembles the shape of a chemically unrelated molecule. In this case, the three-dimensional structure of release factors (made entirely of protein) resembles the shape and charge distribution of a tRNA molecule (**Figure 6–75**). This shape and charge mimicry helps them enter the A-site on the ribosome and cause translation termination.

During translation, the nascent polypeptide moves through a large, water-filled tunnel (approximately 10 nm × 1.5 nm) in the large subunit of the ribosome (see Figure 6–70C). The walls of this tunnel, made primarily of 23S rRNA, are a patchwork of tiny hydrophobic surfaces embedded in a more extensive hydrophilic surface. This structure is not complementary to any peptide, and thus provides a "Teflon" coating through which a polypeptide chain can easily slide. The dimensions of the tunnel suggest that nascent proteins are largely unstructured as they pass through the ribosome, although some α-helical regions of the protein can form before leaving the ribosome tunnel. As it leaves the ribosome, a newly synthesized protein must fold into its proper three-dimensional conformation to be useful to the cell, and later in this chapter we discuss how this folding occurs. First, however, we describe several additional aspects of the translation process itself.

## Proteins Are Made on Polyribosomes

The synthesis of most protein molecules takes between 20 seconds and several minutes. During this very short period, however, it is usual for multiple initiations to take place on each mRNA molecule being translated. As soon as the preceding ribosome has translated enough of the nucleotide sequence to move out

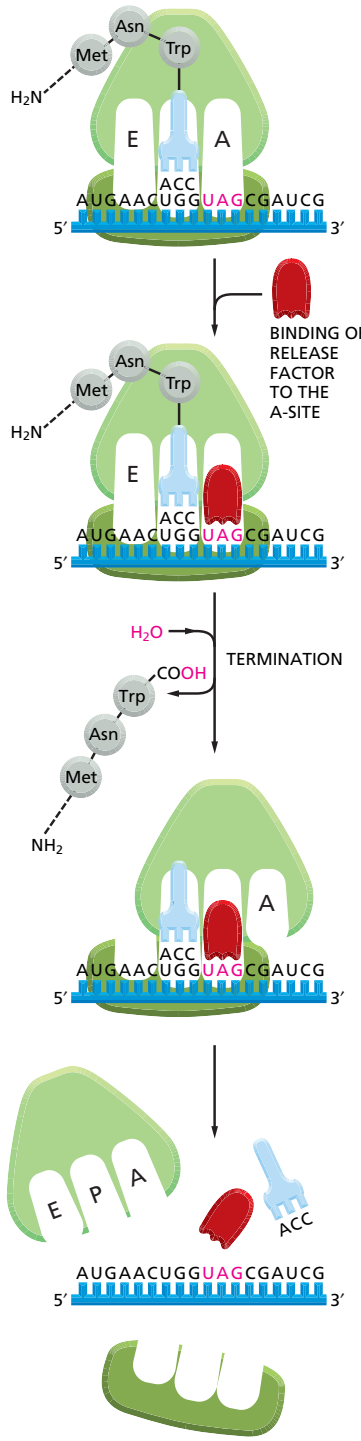



**Figure 6–74 The final phase of protein synthesis.** The binding of a release factor to an A-site bearing a stop codon terminates translation. The completed polypeptide is released and, in a series of reactions that requires additional proteins and GTP hydrolysis (not shown), the ribosome dissociates into its two separate subunits.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 128 of 240



**Figure 6–75 The structure of a human translation release factor (eRF1) and its resemblance to a tRNA molecule.** The protein is on the *left* and the tRNA on the *right.* (From H. Song et al., *Cell* 100:311–321, 2000. With permission from Elsevier.)

of the way, the 5′ end of the mRNA is threaded into a new ribosome. The mRNA molecules being translated are therefore usually found in the form of *polyribosomes* (or *polysomes*): large cytoplasmic assemblies made up of several ribosomes spaced as close as 80 nucleotides apart along a single mRNA molecule (**Figure 6–76**). These multiple initiations allow the cell to make many more protein molecules in a given time than would be possible if each had to be completed before the next could start. <GAAG>

Both bacteria and eucaryotes use polysomes, and both employ additional strategies to speed up the overall rate of protein synthesis even further. Because bacterial mRNA does not need to be processed and is accessible to ribosomes while it is being made, ribosomes can attach to the free end of a bacterial mRNA molecule and start translating it even before the transcription of that RNA is complete, following closely behind the RNA polymerase as it moves along DNA. In eucaryotes, as we have seen, the 5′ and 3′ ends of the mRNA interact (see Figures 6–40 and 6–76A); therefore, as soon as a ribosome dissociates, its two subunits are in an optimal position to reinitiate translation on the same mRNA molecule.

## There Are Minor Variations in the Standard Genetic Code

As discussed in Chapter 1, the genetic code (shown in Figure 6–50) applies to all three major branches of life, providing important evidence for the common

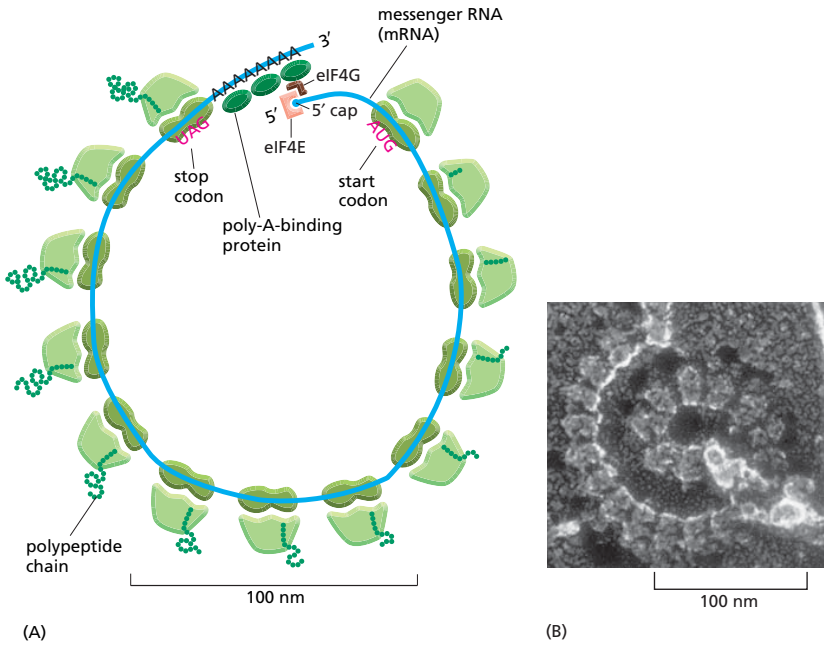



**Figure 6–76 A polyribosome.** (A) Schematic drawing showing how a series of ribosomes can simultaneously translate the same eucaryotic mRNA molecule. (B) Electron micrograph of a polyribosome from a eucaryotic cell. (B, courtesy of John Heuser.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 129 of 240

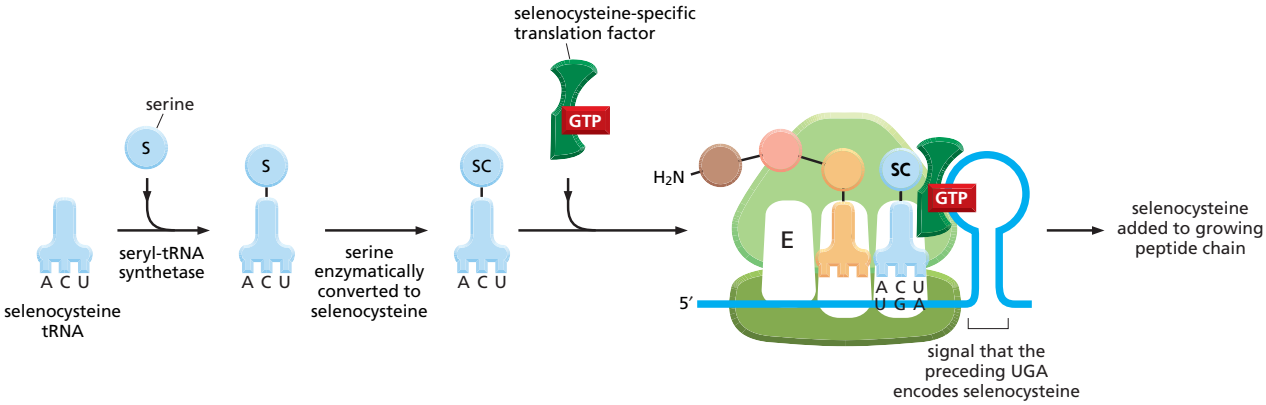



**Figure 6–77 Incorporation of selenocysteine into a growing polypeptide chain.** A specialized tRNA is charged with serine by the normal seryl-tRNA synthetase, and the serine is subsequently converted enzymatically to selenocysteine. A specific RNA structure in the mRNA (a stem and loop structure with a particular nucleotide sequence) signals that selenocysteine is to be inserted at the neighboring UGA codon. As indicated, this event requires the participation of a selenocysteine-specific translation factor.

ancestry of all life on Earth. Although rare, there are exceptions to this code. For example, *Candida albicans,* the most prevalent human fungal pathogen, translates the codon CUG as serine, whereas nearly all other organisms translate it as leucine. Mitochondria (which have their own genomes and encode much of their translational apparatus) often deviate from the standard code. For example, in mammalian mitochondria AUA is translated as methionine, whereas in the cytosol of the cell it is translated as isoleucine (see Table 14–3, p. 862). This type of deviation in the genetic code is "hardwired" into the organisms or the organelles in which it occurs.

A different type of variation, sometimes called *translation recoding*, occurs in many cells. In this case, other nucleotide sequence information present in an mRNA can change the meaning of the genetic code at a particular site in the mRNA molecule. The standard code allows cells to manufacture proteins using only 20 amino acids. However, bacteria, archaea, and eucaryotes have available to them a twenty-first amino acid that can be incorporated directly into a growing polypeptide chain through translation recoding. Selenocysteine, which is essential for the efficient function of a variety of enzymes, contains a selenium atom in place of the sulfur atom of cysteine. Selenocysteine is enzymatically produced from a serine attached to a special tRNA molecule that base-pairs with the UGA codon, a codon normally used to signal a translation stop. The mRNAs for proteins in which selenocysteine is to be inserted at a UGA codon carry an additional nucleotide sequence in the mRNA nearby that causes this recoding event (**Figure 6–77**).

Another form of recoding, *translational frameshifting*, allows more than one protein to be synthesized from a single mRNA. Retroviruses, members of a large group of eucaryotic-infecting pathogens, commonly use translational frameshifting to make both the capsid proteins *(Gag proteins)* and the viral reverse transcriptase and integrase *(Pol proteins)* from the same RNA transcript (see Figure 5–73). The virus needs many more copies of the Gag proteins than it does of the Pol proteins. This quantitative adjustment is achieved by encoding the *Pol* genes just after the *Gag* genes but in a different reading frame. Small amounts of the Pol gene products are made because, on occasion, an upstream translational frameshift allows the Gag protein stop codon to be bypassed. This frameshift occurs at a particular codon in the mRNA and requires a specific *recoding signal,* which seems to be a structural feature of the RNA sequence downstream of this site (**Figure 6–78**).

## Inhibitors of Procaryotic Protein Synthesis Are Useful as Antibiotics

Many of the most effective antibiotics used in modern medicine are compounds made by fungi that inhibit bacterial protein synthesis. Fungi and bacteria compete for many of the same environmental niches, and millions of years of coevolution has resulted in fungi producing potent bacterial inhibitors. Some of these

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 130 of 240

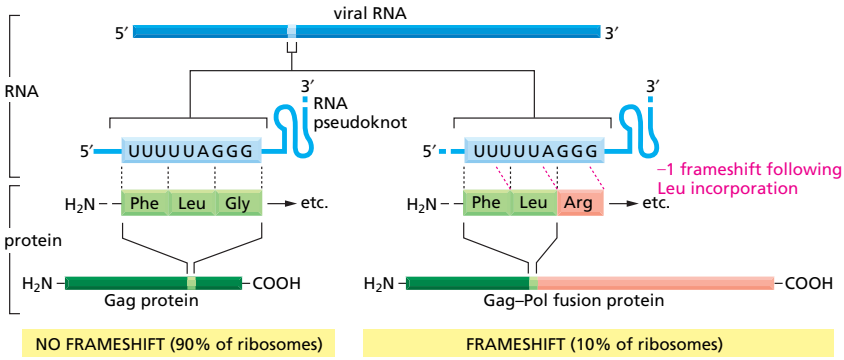



**Figure 6–78 The translational frameshifting that produces the reverse transcriptase and integrase of a retrovirus.** The viral reverse transcriptase and integrase are produced by proteolytic processing of a large protein (the Gag–Pol fusion protein) consisting of both the Gag and Pol amino acid sequences. Proteolytic processing of the more abundant Gag protein produces the viral capsid proteins. Both the Gag and the Gag–Pol fusion proteins start with identical mRNA, but whereas the Gag protein terminates at a stop codon downstream of the sequence shown, translation of the Gag–Pol fusion protein bypasses this stop codon, allowing the synthesis of the longer Gag–Pol fusion protein. The stop-codon-bypass is made possible by a controlled translational frameshift, as illustrated. Features in the local RNA structure (including the RNA loop shown) cause the $tRNA^{Leu}$ attached to the C-terminus of the growing polypeptide chain occasionally to slip backward by one nucleotide on the ribosome, so that it pairs with a UUU codon instead of the UUA codon that had initially specified its incorporation; the next codon (AGG) in the new reading frame specifies an arginine rather than a glycine. This controlled slippage is due in part to a *pseudoknot* that forms in the viral mRNA (see Figure 6–102). The sequence shown is from the human AIDS virus, HIV. (Adapted from T. Jacks et al., *Nature* 331:280–283, 1988. With permission from Macmillan Publishers Ltd.)

drugs exploit the structural and functional differences between bacterial and eucaryotic ribosomes so as to interfere preferentially with the function of bacterial ribosomes. Thus humans can take high dosages of some of these compounds without undue toxicity. Many antibiotics lodge in pockets in the ribosomal RNAs and simply interfere with the smooth operation of the ribosome (**Figure 6–79**). **Table 6–4** lists some of the more common antibiotics of this kind along with several other inhibitors of protein synthesis, some of which act on eucaryotic cells and therefore cannot be used as antibiotics.

Because they block specific steps in the processes that lead from DNA to protein, many of the compounds listed in Table 6–4 are useful for cell biological studies. Among the most commonly used drugs in such investigations are *chloramphenicol*, *cycloheximide*, and *puromycin*, all of which specifically inhibit protein synthesis. In a eucaryotic cell, for example, chloramphenicol inhibits protein synthesis on ribosomes only in mitochondria (and in chloroplasts in plants), presumably reflecting the procaryotic origins of these organelles (discussed in Chapter 14). Cycloheximide, in contrast, affects only ribosomes in the cytosol. Puromycin is especially interesting because it is a structural analog of a tRNA molecule linked to an amino acid and is therefore another example of molecular mimicry; the ribosome mistakes it for an authentic amino acid and covalently incorporates it at the C-terminus of the growing peptide chain, thereby causing the premature termination and release of the polypeptide. As might be expected, puromycin inhibits protein synthesis in both procaryotes and eucaryotes.

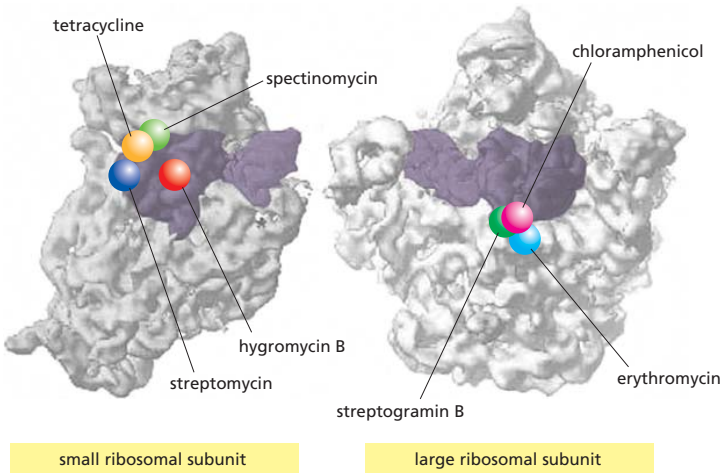



**Figure 6–79 Binding sites for antibiotics on the bacterial ribosome.** The small *(left)* and large *(right)* subunits of the ribosome are arranged as though the ribosome has been opened like a book; the bound tRNA molecules are shown in *purple* (see Figure 6–64). Most of the antibiotics shown bind directly to pockets formed by the ribosomal RNA molecules. Hygromycin B induces errors in translation, spectinomycin blocks the translocation of the peptidyl-tRNA from the A-site to the P-site, and streptogramin B prevents elongation of nascent peptides. Table 6–4 lists the inhibitory mechanisms of the other antibiotics shown in the figure. (Adapted from J. Poehlsgaard and S. Douthwaite, *Nat. Rev. Microbiol.* 3:870–881, 2005. With permission from Macmillan Publishers Ltd.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 131 of 240

**Table 6–4 Inhibitors of Protein or RNA Synthesis**

| INHIBITOR | SPECIFIC EFFECT |
|---|---|
| *Acting only on bacteria* | |
| Tetracycline | blocks binding of aminoacyl-tRNA to A-site of ribosome |
| Streptomycin | prevents the transition from translation initiation to chain elongation and also causes miscoding |
| Chloramphenicol | blocks the peptidyl transferase reaction on ribosomes (step 2 in Figure 6–66) |
| Erythromycin | binds in the exit channel of the ribosome and thereby inhibits elongation of the peptide chain |
| Rifamycin | blocks initiation of RNA chains by binding to RNA polymerase (prevents RNA synthesis) |
| *Acting on bacteria and eucaryotes* | |
| Puromycin | causes the premature release of nascent polypeptide chains by its addition to the growing chain end |
| Actinomycin D | binds to DNA and blocks the movement of RNA polymerase (prevents RNA synthesis) |
| *Acting on eucaryotes but not bacteria* | |
| Cycloheximide | blocks the translocation reaction on ribosomes (step 3 in Figure 6–66) |
| Anisomycin | blocks the peptidyl transferase reaction on ribosomes (step 2 in Figure 6–66) |
| α-Amanitin | blocks mRNA synthesis by binding preferentially to RNA polymerase II |

The ribosomes of eucaryotic mitochondria (and chloroplasts) often resemble those of bacteria in their sensitivity to inhibitors. Therefore, some of these antibiotics can have a deleterious effect on human mitochondria.

## Accuracy in Translation Requires the Expenditure of Free Energy

Translation by the ribosome is a compromise between the opposing constraints of accuracy and speed. We have seen, for example, that the accuracy of translation (1 mistake per $10^4$ amino acids joined) requires time delays each time a new amino acid is added to a growing polypeptide chain, producing an overall speed of translation of 20 amino acids incorporated per second in bacteria. Mutant bacteria with a specific alteration in the small ribosomal subunit have longer delays and translate mRNA into protein with an accuracy considerably higher than this; however, protein synthesis is so slow in these mutants that the bacteria are barely able to survive.

We have also seen that attaining the observed accuracy of protein synthesis requires the expenditure of a great deal of free energy; this is expected, since, as discussed in Chapter 2, there is a price to be paid for any increase in order in the cell. In most cells, protein synthesis consumes more energy than any other biosynthetic process. At least four high-energy phosphate bonds are split to make each new peptide bond: two are consumed in charging a tRNA molecule with an amino acid (see Figure 6–56), and two more drive steps in the cycle of reactions occurring on the ribosome during synthesis itself (see Figure 6–67). In addition, extra energy is consumed each time that an incorrect amino acid linkage is hydrolyzed by a tRNA synthetase (see Figure 6–59) and each time that an incorrect tRNA enters the ribosome, triggers GTP hydrolysis, and is rejected (see Figure 6–67). To be effective, these proofreading mechanisms must also allow an appreciable fraction of correct interactions to be removed; for this reason, proofreading is even more costly in energy than it might seem.

## Quality Control Mechanisms Act to Prevent Translation of Damaged mRNAs

In eucaryotes, mRNA production involves both transcription and a series of elaborate RNA-processing steps; these take place in the nucleus, segregated from ribosomes, and only when the processing is complete are the mRNAs transported to the cytoplasm to be translated (see Figure 6–40). However, this scheme is not foolproof, and some incorrectly processed mRNAs are inadvertently sent to the cytoplasm. In addition, mRNAs that were flawless when they left the nucleus can become broken or otherwise damaged in the cytosol. The danger of translating damaged or incompletely processed mRNAs (which would produce truncated or otherwise aberrant proteins) is apparently so great that the cell has several backup measures to prevent this from happening.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 132 of 240

To avoid translating broken mRNAs, the 5′ cap and the poly-A tail are both recognized by the translation-initiation machinery before translation begins (see Figure 6–72). To help ensure that mRNAs are properly spliced before they are translated, the exon junction complex (EJC), which is deposited on the mRNA following splicing (see Figure 6–40), stimulates the subsequent translation of the mRNA.

But the most powerful mRNA surveillance system, called **nonsense-mediated mRNA decay**, eliminates defective mRNAs before they can be efficiently translated into protein. This mechanism is brought into play when the cell determines that an mRNA molecule has a nonsense (stop) codon (UAA, UAG, or UGA) in the "wrong" place—a situation likely to arise in an mRNA molecule that has been improperly spliced. Aberrant splicing will usually result in the random introduction of a nonsense codon into the reading frame of the mRNA, especially in organisms, such as humans, that have a large average intron size (see Figure 6–32B).

This surveillance mechanism begins as an mRNA molecule is being transported from the nucleus to the cytosol. As its 5′ end emerges from the nuclear pore, the mRNA is met by a ribosome, which begins to translate it. As translation proceeds, the exon junction complexes (EJC) bound to the mRNA at each splice-site are apparently displaced by the moving ribosome. The normal stop codon will be within the last exon, so by the time the ribosome reaches it and stalls, no more EJCs should be bound to the mRNA. If this is the case, the mRNA "passes inspection" and is released to the cytosol where it can be translated in earnest (**Figure 6–80**). However, if the ribosome reaches a premature stop codon and stalls, it senses that EJCs remain and the bound mRNA molecule is rapidly degraded. In this way, the first round of translation allows the cell to test the fitness of each mRNA molecule as it exits the nucleus.

Nonsense-mediated decay may have been especially important in evolution, allowing eucaryotic cells to more easily explore new genes formed by DNA rearrangements, mutations, or alternative patterns of splicing—by selecting only those mRNAs for translation that can produce a full-length protein. Nonsense-mediated decay is also important in cells of the developing immune system, where the extensive DNA rearrangements that occur (see Figure 25–36) often generate premature termination codons. The surveillance system degrades the mRNAs produced from such rearranged genes, thereby avoiding the potential toxic effects of truncated proteins.

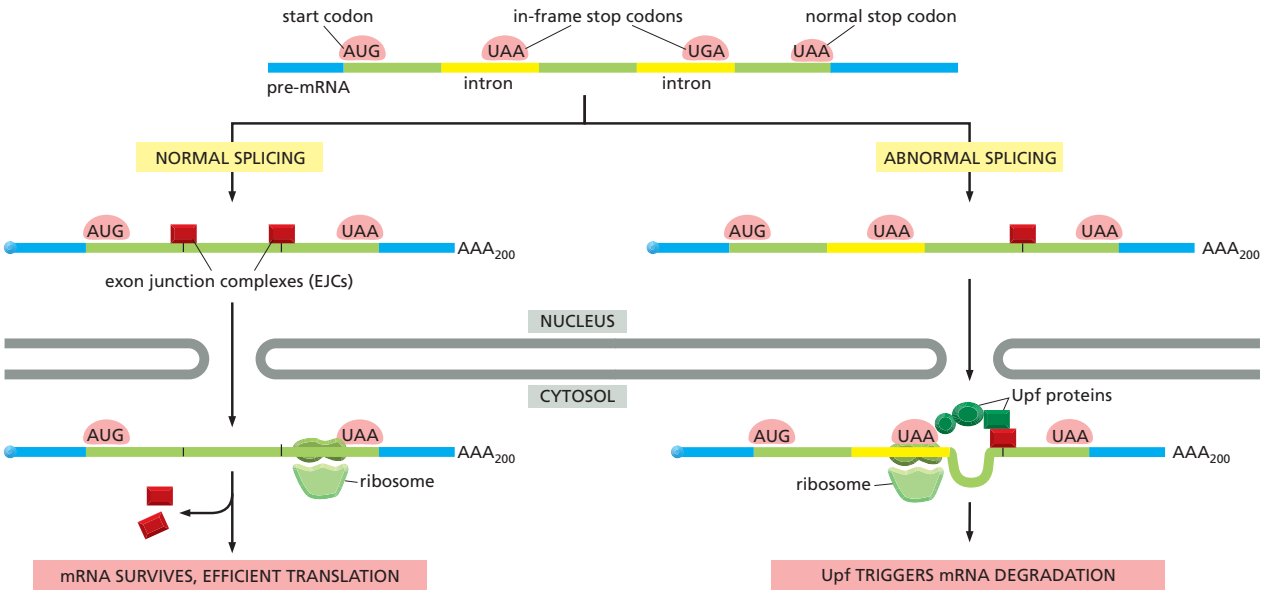



**Figure 6–80 Nonsense-mediated mRNA decay.** As shown on the right, the failure to correctly splice a pre-mRNA often introduces a premature stop codon into the reading frame for the protein. The introduction of such an "in-frame" stop codon is particularly likely to occur in mammals, where the introns tend to be very long. When translated, these abnormal mRNAs produce aberrant proteins, which could damage the cell. However, as shown at the bottom right of the figure, these abnormal RNAs are destroyed by the nonsense-mediated decay mechanism. According to one model, an mRNA molecule, bearing exon junction complexes (EJCs) to mark successfully completed splices, is first met by a ribosome that performs a "test" round of translation. As the mRNA passes through the tight channel of the ribosome, the EJCs are stripped off, and successful mRNAs are released to undergo multiple rounds of translation *(left side)*. However, if an in-frame stop codon is encountered before the final exon junction complex is reached (*right side*), the mRNA undergoes nonsense-mediated decay, which is triggered by the Upf proteins (*green*) that bind to each EJC. Note that, to trigger nonsense-mediated decay, the premature stop codon must be in the same reading frame as that of the normal protein. (Adapted from J. Lykke-Andersen et al., *Cell* 103:1121–1131, 2000. With permission from Elsevier.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 133 of 240

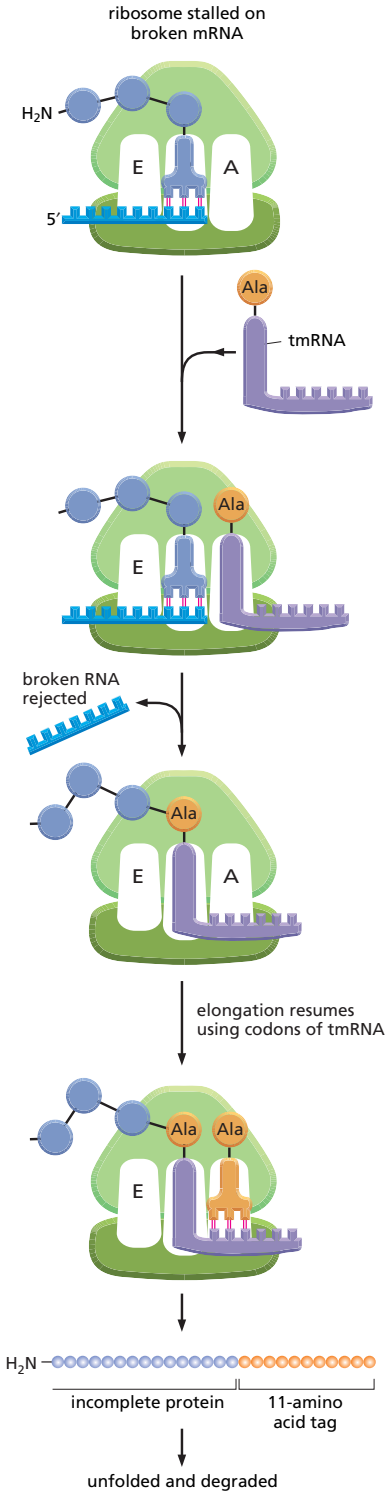



**Figure 6–81 The rescue of a bacterial ribosome stalled on an incomplete mRNA molecule.** The tmRNA shown is a 363-nucleotide RNA with both tRNA and mRNA functions, hence its name. It carries an alanine and can enter the vacant A-site of a stalled ribosome to add this alanine to a polypeptide chain, mimicking a tRNA although no codon is present to guide it. The ribosome then translates 10 codons from the tmRNA, completing an 11 amino acid tag on the protein. Proteases recognize this tag and degrade the entire protein. Although the example shown in the figure is from bacteria, eucaryotes can employ a similar strategy.

Finally, the nonsense-mediated surveillance pathway plays an important role in mitigating the symptoms of many inherited human diseases. As we have seen, inherited diseases are usually caused by mutations that spoil the function of a key protein, such as hemoglobin or one of the blood clotting factors. Approximately one-third of all genetic disorders in humans result from nonsense mutations or mutations (such as frameshift mutations or splice-site mutations) that place nonsense mutations into the gene's reading frame. In individuals that carry one mutant and one functional gene, nonsense-mediated decay eliminates the aberrant mRNA and thereby prevents a potentially toxic protein from being made. Without this safeguard, individuals with one functional and one mutant "disease gene" would likely suffer much more severe symptoms.

We saw earlier in this chapter that bacteria lack the elaborate mRNA processing found in eucaryotes and that translation often begins before the synthesis of the RNA molecule is completed. Yet bacteria also have quality control mechanisms to deal with incompletely synthesized and broken mRNAs. When the bacterial ribosome translates to the end of an incomplete RNA it stalls and does not release the RNA. Rescue comes in the form of a special RNA (called tmRNA), which enters the A-site of the ribosome and is itself translated, releasing the ribosome. The special 11 amino acid tag thus added to the C-terminus of the truncated protein signals to proteases that the entire protein is to be degraded (**Figure 6–81**).

## Some Proteins Begin to Fold While Still Being Synthesized

The process of gene expression is not over when the genetic code has been used to create the sequence of amino acids that constitutes a protein. To be useful to the cell, this new polypeptide chain must fold up into its unique three-dimensional conformation, bind any small-molecule cofactors required for its activity, be appropriately modified by protein kinases or other protein-modifying enzymes, and assemble correctly with the other protein subunits with which it functions (**Figure 6–82**).

The information needed for all of the steps listed above is ultimately contained in the sequence of linked amino acids that the ribosome produces when it translates an mRNA molecule into a polypeptide chain. As discussed in Chapter 3, when a protein folds into a compact structure, it buries most of its hydrophobic residues in an interior core. In addition, large numbers of noncovalent interactions form between various parts of the molecule. It is the sum of all of these energetically favorable arrangements that determines the final folding pattern of the polypeptide chain—as the conformation of lowest free energy (see p. 130).

Through many millions of years of evolution, the amino acid sequence of each protein has been selected not only for the conformation that it adopts but also for an ability to fold rapidly. For some proteins, this folding begins immediately, as the protein spins out of the ribosome, starting from the N-terminal end. In these cases, as each protein domain emerges from the ribosome, within a few seconds it forms a compact structure that contains most of the final secondary features (α helices and β sheets) aligned in roughly the right conformation (**Figure 6–83**). For many protein domains, this unusually dynamic and flexible state called a *molten globule*, is the starting point for a relatively slow process in which many side-chain adjustments occur that eventually form the correct tertiary

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 134 of 240

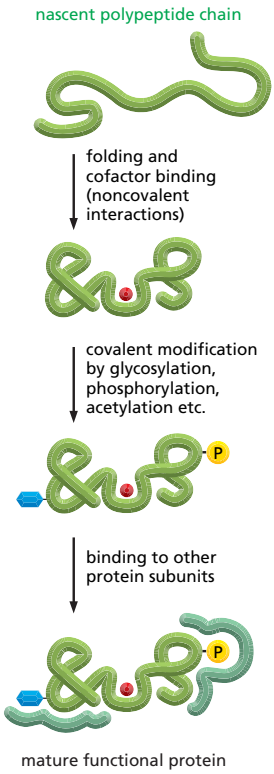



**Figure 6–82 Steps in the creation of a functional protein.** As indicated, translation of an mRNA sequence into an amino acid sequence on the ribosome is not the end of the process of forming a protein. To function, the completed polypeptide chain must fold correctly into its three-dimensional conformation, bind any cofactors required, and assemble with its partner protein chains (if any). Noncovalent bond formation drives these changes. As indicated, many proteins also require covalent modifications of selected amino acids. Although the most frequent modifications are protein glycosylation and protein phosphorylation, more than 100 different types of covalent modifications are known (see, for example, Figure 3–81).

structure. It takes several minutes to synthesize a protein of average size, and for some proteins much of the folding process is complete by the time the ribosome releases the C-terminal end of a protein (**Figure 6–84**).

## Molecular Chaperones Help Guide the Folding of Most Proteins

Most proteins probably do not begin to fold during their synthesis. Instead, they are met at the ribosome by a special class of proteins called **molecular chaperones**. Molecular chaperones are useful for cells because there are many different paths that can be taken to convert an unfolded or partially folded protein to its final compact conformation. For many proteins, some of the intermediates formed along the way would aggregate and be left as off-pathway dead ends without the intervention of a chaperone (**Figure 6–85**).

Many molecular chaperones are called *heat-shock proteins* (designated *Hsp*), because they are synthesized in dramatically increased amounts after a brief exposure of cells to an elevated temperature (for example, 42°C for cells that normally live at 37°C). This reflects the operation of a feedback system that responds to an increase in misfolded proteins (such as those produced by elevated temperatures) by boosting the synthesis of the chaperones that help these proteins refold.

There are several major families of eucaryotic molecular chaperones, including the Hsp60 and Hsp70 proteins. Different family members function in different organelles. Thus, as discussed in Chapter 12, mitochondria contain their own Hsp60 and Hsp70 molecules that are distinct from those that function in the cytosol; and a special Hsp70 (called *BIP*) helps to fold proteins in the endoplasmic reticulum.

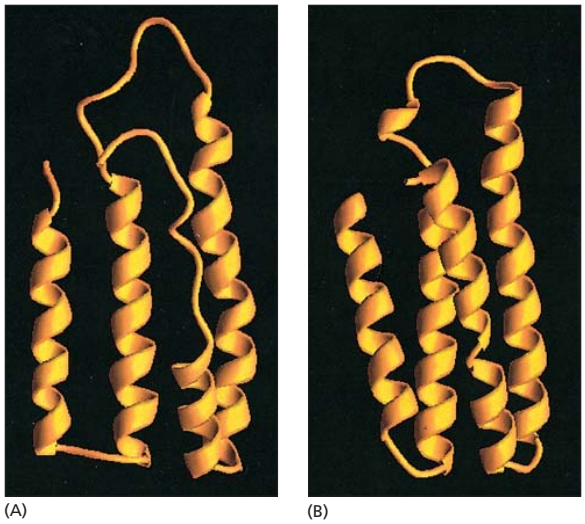



**Figure 6–83 The structure of a molten globule.** (A) A molten globule form of cytochrome $b_{562}$ is more open and less highly ordered than the final folded form of the protein, shown in (B). Note that the molten globule contains most of the secondary structure of the final form, although the ends of the α helices are unravelled and one of the helices is only partly formed. (Courtesy of Joshua Wand, from Y. Feng et al., *Nat. Struct. Biol.* 1:30–35, 1994. With permission from Macmillan Publishers Ltd.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 135 of 240

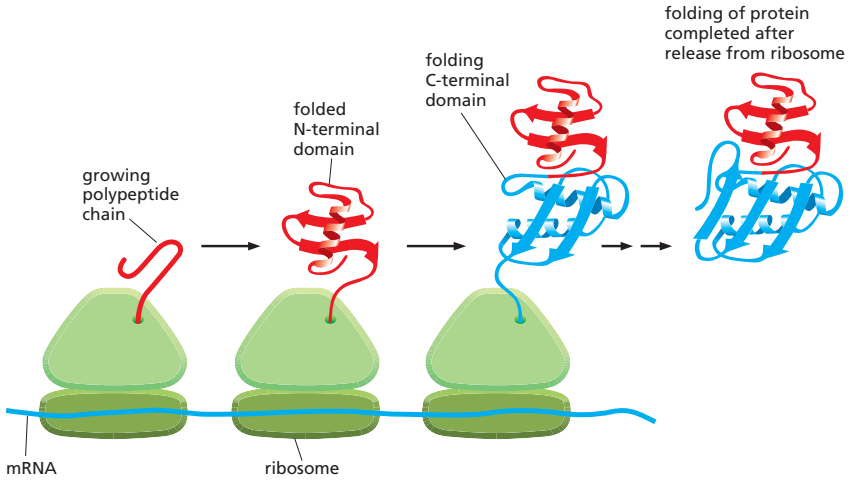



**Figure 6–84 Co-translational protein folding.** A growing polypeptide chain is shown acquiring its secondary and tertiary structure as it emerges from a ribosome. The N-terminal domain folds first, while the C-terminal domain is still being synthesized. This protein has not achieved its final conformation at the time it is released from the ribosome. (Modified from A.N. Federov and T.O. Baldwin, *J. Biol. Chem.* 272:32715–32718, 1997.)

The Hsp60 and Hsp70 proteins each work with their own small set of associated proteins when they help other proteins to fold. Hsps share an affinity for the exposed hydrophobic patches on incompletely folded proteins, and they hydrolyze ATP, often binding and releasing their protein substrate with each cycle of ATP hydrolysis. In other respects, the two types of Hsp proteins function differently. The Hsp70 machinery acts early in the life of many proteins, binding to a string of about seven hydrophobic amino acids before the protein leaves the ribosome (**Figure 6–86**). In contrast, Hsp60-like proteins form a large barrel-shaped structure that acts after a protein has been fully synthesized. This type of chaperone, sometimes called a *chaperonin*, forms an "isolation chamber" into which misfolded proteins are fed, preventing their aggregation and providing them with a favorable environment in which to attempt to refold (**Figure 6–87**).

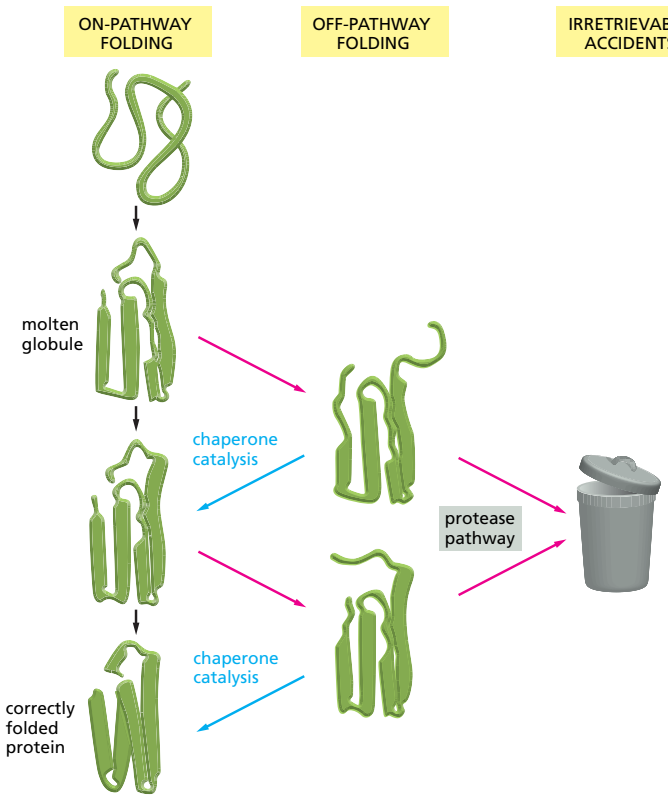



**Figure 6–85 A current view of protein folding.** Each domain of a newly synthesized protein rapidly attains a "molten globule" state. Subsequent folding occurs more slowly and by multiple pathways, often involving the help of a molecular chaperone. Some molecules may still fail to fold correctly; as explained in the text, specific proteases recognize and degrade these molecules.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 136 of 240

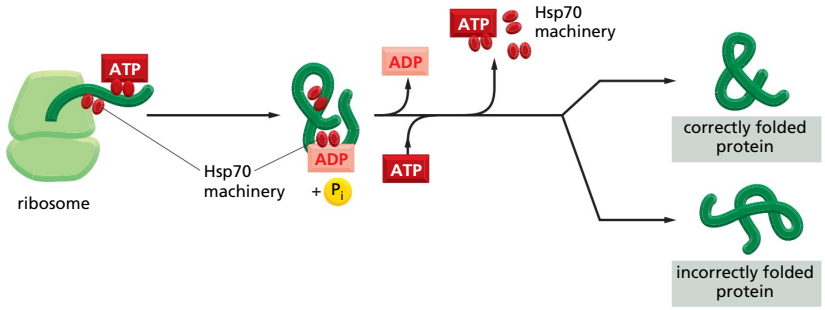



**Figure 6–86 The Hsp70 family of molecular chaperones.** These proteins act early, recognizing a small stretch of hydrophobic amino acids on a protein's surface. Aided by a set of smaller Hsp40 proteins (not shown), ATP-bound Hsp70 molecules grasp their target protein and then hydrolyze ATP to ADP, undergoing conformational changes that cause the Hsp70 molecules to associate even more tightly with the target. After the Hsp40 dissociates, the rapid rebinding of ATP induces the dissociation of the Hsp70 protein after ADP release. In reality, repeated cycles of Hsp protein binding and release help the target protein to refold, as schematically illustrated in Figure 6–85.

The chaperones shown in Figures 6–86 and 6–87 often use many cycles of ATP hydrolysis to fold a single polypeptide chain correctly. Although some of this energy expenditure is used to perform mechanical work, probably much more is expended to ensure that protein folding is accurate. Just as we saw for transcription, splicing, and translation, the expenditure of free energy can be used by cells to improve the accuracy of a biological process. In the case of protein folding, ATP hydrolysis allows chaperones to recognize a wide variety of misfolded structures, to halt any further misfolding and to recommence folding of a protein in an orderly way.

Although our discussion focuses on only two types of chaperones, the cell has a variety of others. The enormous diversity of proteins in cells presumably requires a wide range of chaperones with versatile surveillance and correction capabilities.

## Exposed Hydrophobic Regions Provide Critical Signals for Protein Quality Control

If radioactive amino acids are added to cells for a brief period, the newly synthesized proteins can be followed as they mature into their final functional form.

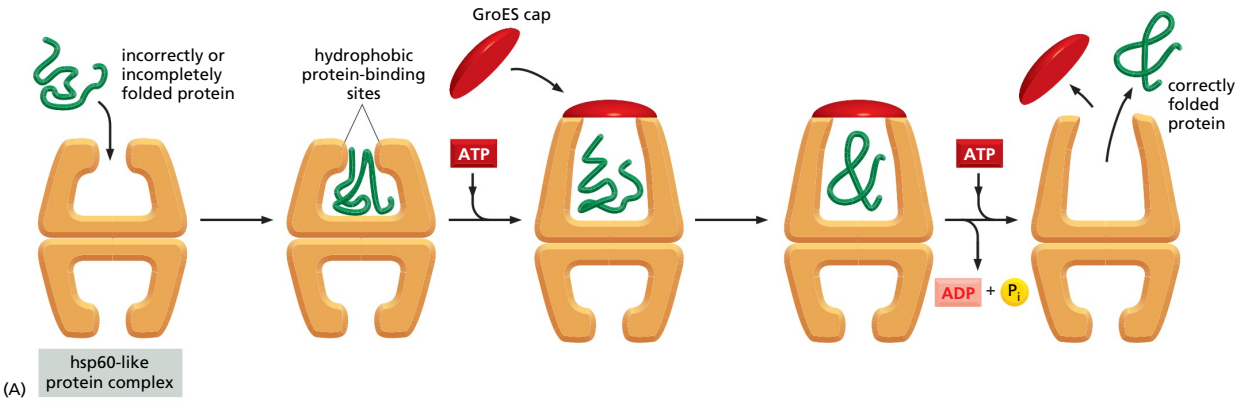
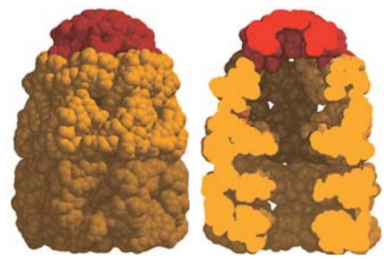



**Figure 6–87 The structure and function of the Hsp60 family of molecular chaperones.** (A) The catalysis of protein refolding. A misfolded protein is initially captured by hydrophobic interactions along one rim of the barrel. The subsequent binding of ATP plus a protein cap increases the diameter of the barrel rim, which may transiently stretch (partly unfold) the client protein. This also confines the protein in an enclosed space, where it has a new opportunity to fold. After about 15 seconds, ATP hydrolysis occurs, weakening the complex. Subsequent binding of another ATP molecule ejects the protein, whether folded or not, and the cycle repeats. This type of molecular chaperone is also known as a chaperonin; it is designated as Hsp60 in mitochondria, TCP1 in the cytosol of vertebrate cells, and GroEL in bacteria. As indicated, only half of the symmetrical barrel operates on a client protein at any one time. (B) The structure of GroEL bound to its GroES cap, as determined by X-ray crystallography. On the *left* is shown the outside of the barrel-like structure and on the *right* a cross section through its center. (B, adapted from B. Bukau and A.L. Horwich, *Cell* 92:351–366, 1998. With permission from Elsevier.)




LCY Biotechnology Holding, Inc.
Ex. 1008
Page 137 of 240

This type of experiment demonstrates that the Hsp70 proteins act first, beginning when a protein is still being synthesized on a ribosome, and the Hsp60-like proteins act only later to help fold completed proteins. But how does the cell distinguish misfolded proteins, which require additional rounds of ATP-catalyzed refolding, from those with correct structures?

Before answering, we need to pause to consider the post-translational fate of proteins more broadly. Usually, if a protein has a sizable exposed patch of hydrophobic amino acids on its surface, it is abnormal: it has either failed to fold correctly after leaving the ribosome, suffered an accident that partly unfolded it at a later time, or failed to find its normal partner subunit in a larger protein complex. Such a protein is not merely useless to the cell, it can be dangerous. Many proteins with an abnormally exposed hydrophobic region can form large aggregates in the cell. We shall see that, in rare cases, such aggregates do form and cause severe human diseases. Normally, however, powerful protein quality control mechanisms prevent such disasters.

Given this background, it is not surprising that cells have evolved elaborate mechanisms that recognize the hydrophobic patches on proteins and minimize the damage they cause. Two of these mechanisms depend on the molecular chaperones just discussed, which bind to the patch and attempt to repair the defective protein by giving it another chance to fold. At the same time, by covering the hydrophobic patches, these chaperones transiently prevent protein aggregation. Proteins that very rapidly fold correctly on their own do not display such patches and the chaperones bypass them.

**Figure 6–88** outlines all of the quality control choices that a cell makes for a difficult-to-fold, newly synthesized protein. As indicated, when attempts to refold a protein fail, a third mechanism is called into play that completely destroys the protein by proteolysis. The proteolytic pathway begins with the recognition of an abnormal hydrophobic patch on a protein's surface, and it ends with the delivery of the entire protein to a protein destruction machine, a complex protease known as the *proteasome*. As described next, this process depends on an elaborate protein-marking system that also carries out other central functions in the cell by destroying selected normal proteins.

## The Proteasome Is a Compartmentalized Protease with Sequestered Active Sites

The proteolytic machinery and the chaperones compete with one another to reorganize a misfolded protein. If a newly synthesized protein folds rapidly, at most only a small fraction of it is degraded. In contrast, a slowly folding protein is vulnerable to the proteolytic machinery for a longer time, and many more of its molecules are destroyed before the remainder attain the proper folded state. Due to mutations or to errors in transcription, RNA splicing, and translation, some proteins never fold properly. It is particularly important that the cell destroy these potentially harmful proteins.

The apparatus that deliberately destroys aberrant proteins is the **proteasome**, an abundant ATP-dependent protease that constitutes nearly 1% of cell protein. Present in many copies dispersed throughout the cytosol and the nucleus, the proteasome also destroys aberrant proteins of the endoplasmic

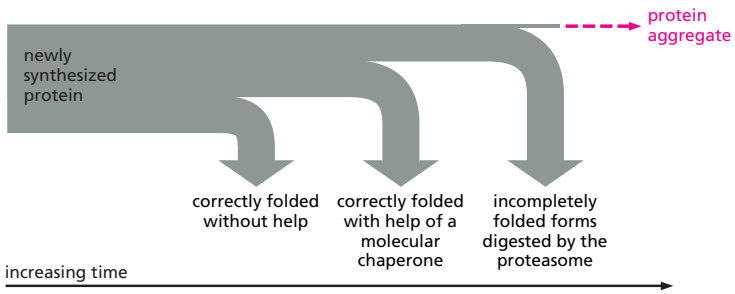



**Figure 6–88 The processes that monitor protein quality following protein synthesis.** A newly synthesized protein sometimes folds correctly and assembles on its own with its partner proteins, in which case the quality control mechanisms leave it alone. Incompletely folded proteins are helped to refold by molecular chaperones: first by a family of Hsp70 proteins, and then in some cases, by Hsp60-like proteins. For both types of chaperones, the client proteins are recognized by an abnormally exposed patch of hydrophobic amino acids on their surface. These "protein-rescue" processes compete with another mechanism that, upon recognizing an abnormally exposed patch, marks the protein for destruction by the proteasome. The combined activity of all of these processes is needed to prevent massive protein aggregation in a cell, which can occur when many hydrophobic regions on proteins clump together nonspecifically.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 138 of 240

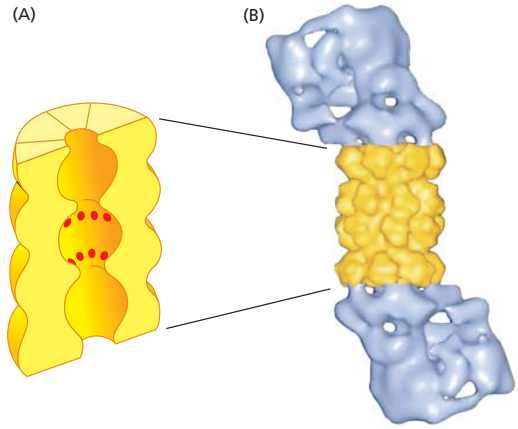



**Figure 6–89 The proteasome.** (A) A cut-away view of the structure of the central 20S cylinder, as determined by x-ray crystallography, with the active sites of the proteases indicated by *red dots*. (B) The entire proteasome, in which the central cylinder *(yellow)* is supplemented by a 19S cap *(blue)* at each end. The cap structure has been determined by computer processing of electron microscope images. The complex cap (also called the regulatory particle) selectively binds proteins that have been marked by ubiquitin for destruction; it then uses ATP hydrolysis to unfold their polypeptide chains and feed them through a narrow channel (see Figure 6–91) into the inner chamber of the 20S cylinder for digestion to short peptides. (B, from W. Baumeister et al., *Cell* 92:367–380, 1998. With permission from Elsevier.)

reticulum (ER). An ER-based surveillance system detects proteins that fail either to fold or to be assembled properly after they enter the ER, and *retrotranslocates* them back to the cytosol for degradation (discussed in Chapter 12).

Each proteasome consists of a central hollow cylinder (the 20S core proteasome) formed from multiple protein subunits that assemble as a quasi-cylindrical stack of four heptameric rings (**Figure 6–89**). Some of the subunits are distinct proteases whose active sites face the cylinder's inner chamber. The design prevents these highly efficient proteases from running rampant through the cell. Each end of the cylinder is normally associated with a large protein complex (the 19S cap), which contains a six-subunit protein ring, through which target proteins are threaded into the proteasome core where they are degraded (**Figure 6–90**). The threading reaction, driven by ATP hydrolysis, unfolds the target proteins as they move through the cap, exposing them to the proteases lining the proteasome core (**Figure 6–91**). The proteins that make up the ring structure in the proteasome cap belong to a large class of protein "unfoldases" known as *AAA proteins*. Many of them function as hexamers, and it is possible that they share mechanistic features with the ATP-dependent unwinding of DNA by DNA helicases (see Figure 5–15).

A crucial property of the proteasome, and one reason for the complexity of its design, is the *processivity* of its mechanism: in contrast to a "simple" protease that cleaves a substrate's polypeptide chain just once before dissociating, the proteasome keeps the entire substrate bound until all of it is converted into short peptides.

The 19S caps also act as regulated "gates" at the entrances to the inner proteolytic chamber, and they are responsible for binding a targeted protein substrate to the proteasome. With a few exceptions, the proteasomes act on proteins that have been specifically marked for destruction by the covalent attachment of a recognition tag formed from a small protein called *ubiquitin* (**Figure 6–92**A). Ubiquitin exists in cells either free or covalently linked to

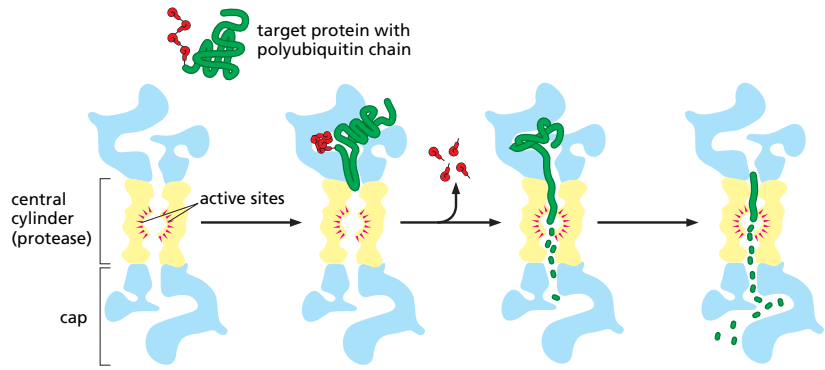



**Figure 6–90 Processive protein digestion by the proteasome.** The proteasome cap recognizes a substrate protein, in this case marked by a polyubiquitin chain (see Figure 6–92), and subsequently translocates it into the proteasome core, where it is digested. At an early stage, the ubiquitin is cleaved from the substrate protein and is recycled. Translocation into the core of the proteasome is mediated by a ring of ATP-dependent proteins that unfold the substrate protein as it is threaded through the ring and into the proteasome core (see Figure 6–91). (From S. Prakash and A. Matouschek, *Trends Biochem. Sci.* 29:593–600, 2004. With permission from Elsevier.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 139 of 240

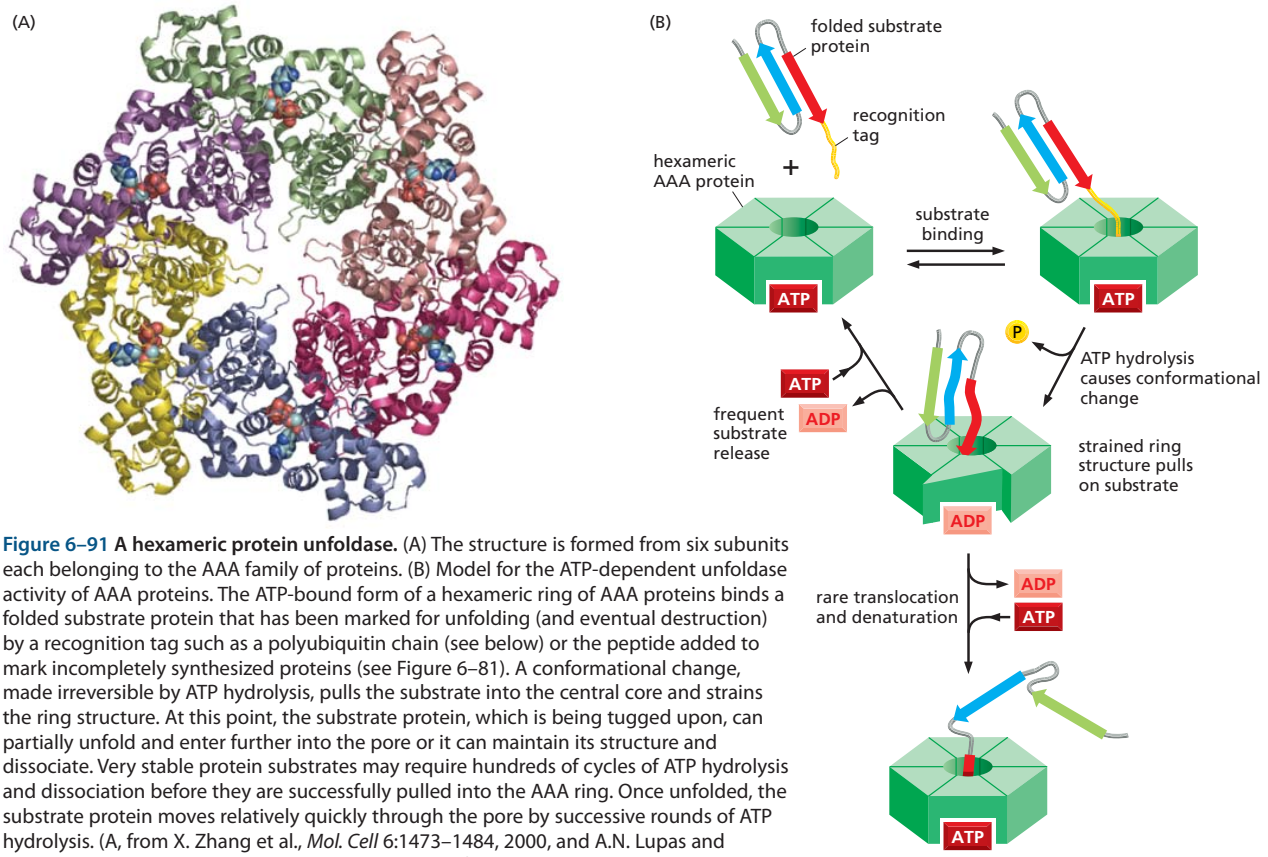



**Figure 6–91 A hexameric protein unfoldase.** (A) The structure is formed from six subunits each belonging to the AAA family of proteins. (B) Model for the ATP-dependent unfoldase activity of AAA proteins. The ATP-bound form of a hexameric ring of AAA proteins binds a folded substrate protein that has been marked for unfolding (and eventual destruction) by a recognition tag such as a polyubiquitin chain (see below) or the peptide added to mark incompletely synthesized proteins (see Figure 6–81). A conformational change, made irreversible by ATP hydrolysis, pulls the substrate into the central core and strains the ring structure. At this point, the substrate protein, which is being tugged upon, can partially unfold and enter further into the pore or it can maintain its structure and dissociate. Very stable protein substrates may require hundreds of cycles of ATP hydrolysis and dissociation before they are successfully pulled into the AAA ring. Once unfolded, the substrate protein moves relatively quickly through the pore by successive rounds of ATP hydrolysis. (A, from X. Zhang et al., *Mol. Cell* 6:1473–1484, 2000, and A.N. Lupas and J. Martin, *Curr. Opin. Struct. Biol.* 12:746–753, 2002; B, from R.T. Sauer et al., *Cell* 119:9–18, 2004. All with permission from Elsevier.)

many different intracellular proteins. For many proteins, tagging by ubiquitin results in their destruction by the proteasome. However, in other cases, ubiquitin tagging has an entirely different meaning. Ultimately, it is the number of ubiquitin molecules added and the way in which they are linked together that determines how the cell interprets the ubiquitin message (**Figure 6–93**). In the following sections, we emphasize the role of ubiquitylation in signifying protein degradation.

## An Elaborate Ubiquitin-Conjugating System Marks Proteins for Destruction

Ubiquitin is prepared for conjugation to other proteins by the ATP-dependent *ubiquitin-activating enzyme* (E1), which creates an activated, E1-bound ubiquitin that is subsequently transferred to one of a set of *ubiquitin-conjugating* (E2) enzymes (Figure 6–92B). The E2 enzymes act in conjunction with accessory (E3) proteins. In the E2–E3 complex, called *ubiquitin ligase*, the E3 component binds to specific degradation signals, called degrons, in protein substrates, helping E2 to form a *polyubiquitin* chain linked to a lysine of the substrate protein. In this chain, the C-terminal residue of each ubiquitin is linked to a specific lysine of the preceding ubiquitin molecule (see Figure 6–93), producing a linear series of ubiquitin–ubiquitin conjugates (Figure 6–92C). It is this polyubiquitin chain on a target protein that is recognized by a specific receptor in the proteasome.

There are roughly 30 structurally similar but distinct E2 enzymes in mammals, and hundreds of different E3 proteins that form complexes with specific E2 enzymes. The ubiquitin–proteasome system thus consists of many distinct but similarly organized proteolytic pathways, which have in common both the

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 140 of 240

E1 enzyme at the "top" and the proteasome at the "bottom," and differ by the compositions of their E2–E3 ubiquitin ligases and accessory factors. Distinct ubiquitin ligases recognize different degradation signals, and therefore target distinct subsets of intracellular proteins for destruction.

Denatured or otherwise misfolded proteins, as well as proteins containing oxidized or other abnormal amino acids, are recognized and destroyed because abnormal proteins tend to present on their surface amino acid sequences or conformational motifs that are recognized as degradation signals by a set of E3 molecules in the ubiquitin–proteasome system; these sequences must of course be buried and therefore inaccessible in the normal counterparts of these proteins. However, a proteolytic pathway that recognizes and destroys abnormal proteins must be able to distinguish between *completed* proteins that have "wrong" conformations and the many growing polypeptides on ribosomes (as well as polypeptides just released from ribosomes) that have not yet achieved their normal folded conformation. This is not a trivial problem; the ubiquitin–proteasome system is thought to destroy many of the nascent and newly formed protein molecules not because these proteins are abnormal as such, but because they transiently expose degradation signals that are buried in their mature (folded) state.

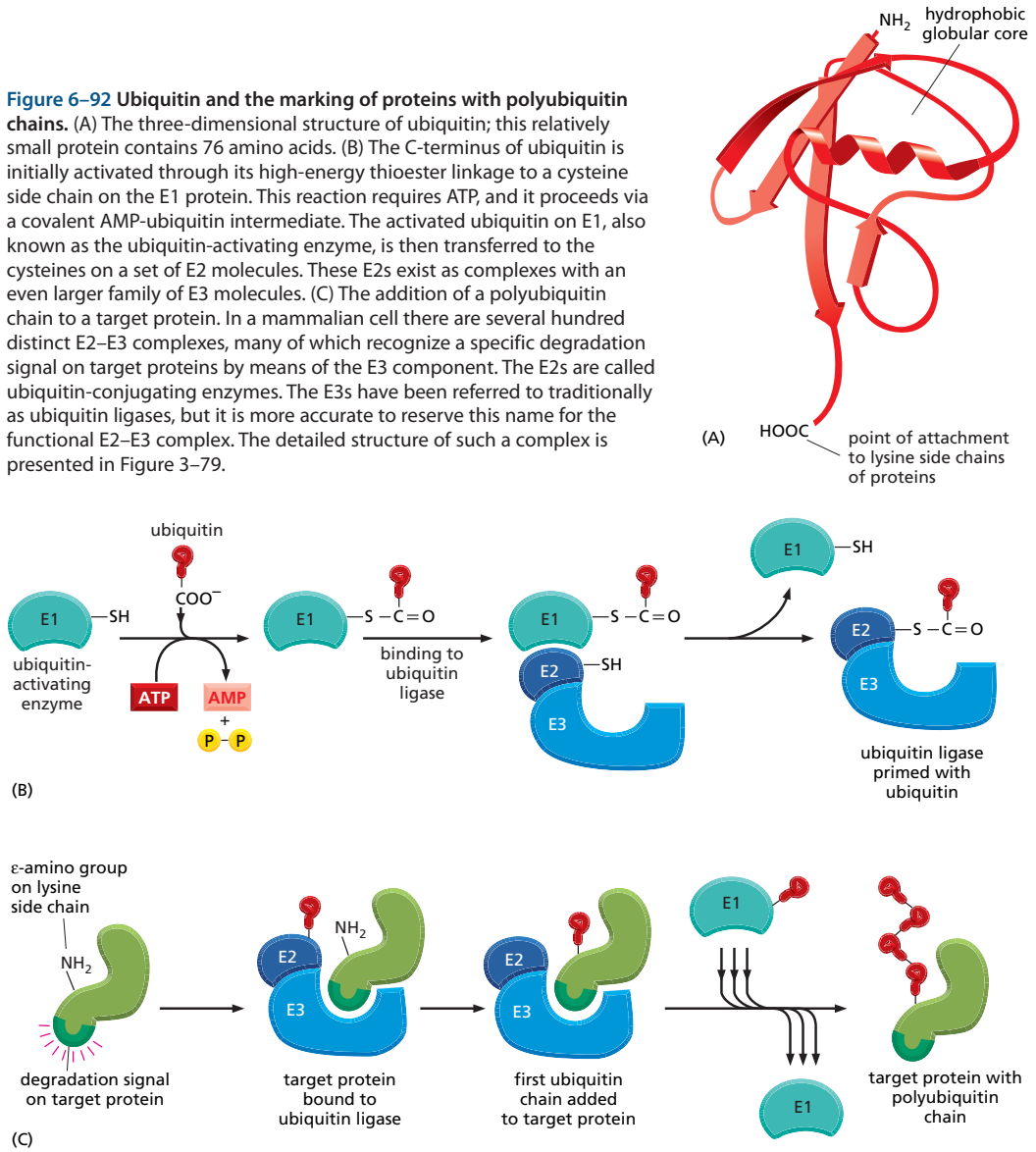



**Figure 6–92 Ubiquitin and the marking of proteins with polyubiquitin chains.** (A) The three-dimensional structure of ubiquitin; this relatively small protein contains 76 amino acids. (B) The C-terminus of ubiquitin is initially activated through its high-energy thioester linkage to a cysteine side chain on the E1 protein. This reaction requires ATP, and it proceeds via a covalent AMP-ubiquitin intermediate. The activated ubiquitin on E1, also known as the ubiquitin-activating enzyme, is then transferred to the cysteines on a set of E2 molecules. These E2s exist as complexes with an even larger family of E3 molecules. (C) The addition of a polyubiquitin chain to a target protein. In a mammalian cell there are several hundred distinct E2–E3 complexes, many of which recognize a specific degradation signal on target proteins by means of the E3 component. The E2s are called ubiquitin-conjugating enzymes. The E3s have been referred to traditionally as ubiquitin ligases, but it is more accurate to reserve this name for the functional E2–E3 complex. The detailed structure of such a complex is presented in Figure 3–79.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 141 of 240

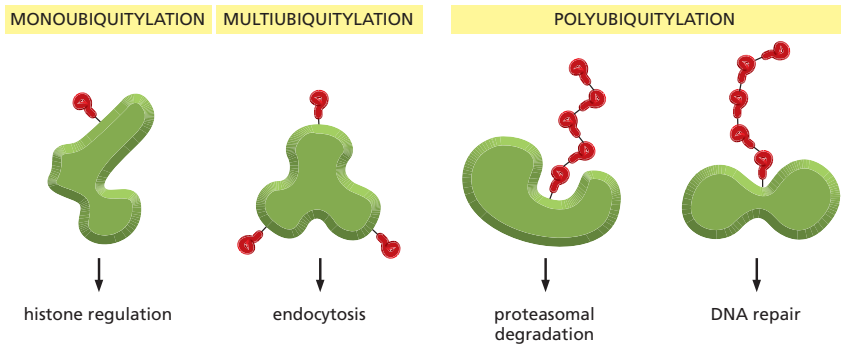



**Figure 6–93 The marking of proteins by ubiquitin.** Each modification pattern shown can have a specific meaning to the cell. The two types of polyubiquitylation differ in the way the ubiquitin molecules are linked together. Linkage through Lys48 signifies degradation by the proteasome whereas that through Lys63 has other meanings. Ubiquitin markings are "read" by proteins that specifically recognize each type of modification.

## Many Proteins Are Controlled by Regulated Destruction

One function of intracellular proteolytic mechanisms is to recognize and eliminate misfolded or otherwise abnormal proteins, as just described. Yet another function of these proteolytic pathways is to confer short lifetimes on specific normal proteins whose concentrations must change promptly with alterations in the state of a cell. Some of these short-lived proteins are degraded rapidly at all times, while many others are *conditionally* short-lived, that is, they are metabolically stable under some conditions but become unstable upon a change in the cell's state. For example, mitotic cyclins are long-lived throughout the cell cycle until their sudden degradation at the end of mitosis, as explained in Chapter 17.

How is such a regulated destruction of a protein controlled? Several mechanisms are illustrated through specific examples that appear later in this book. In one general class of mechanism (**Figure 6–94**A), the activity of a ubiquitin ligase is turned on either by E3 phosphorylation or by an allosteric transition in an E3 protein caused by its binding to a specific small or large molecule. For example, the anaphase-promoting complex (APC) is a multisubunit ubiquitin ligase that is activated by a cell-cycle-timed subunit addition at mitosis. The activated APC then causes the degradation of mitotic cyclins and several other regulators of the metaphase–anaphase transition (see Figure 17–44).

Alternatively, in response either to intracellular signals or to signals from the environment, a degradation signal can be created in a protein, causing its rapid ubiquitylation and destruction by the proteasome. One common way to create such a signal is to phosphorylate a specific site on a protein that unmasks a normally hidden degradation signal. Another way to unmask such a signal is by the regulated dissociation of a protein subunit. Finally, powerful degradation signals can be created by cleaving a single peptide bond, provided that this cleavage creates a new N-terminus that is recognized by a specific E3 as a "destabilizing" N-terminal residue (Figure 6–94B).

The N-terminal type of degradation signal arises because of the "N-end rule," which relates the lifetime of a protein *in vivo* to the identity of its N-terminal residue. There are 12 destabilizing residues in the N-end rule of the yeast *S. cerevisiae* (Arg, Lys, His, Phe, Leu, Tyr, Trp, Ile, Asp, Glu, Asn, and Gln), out of the 20 standard amino acids. The destabilizing N-terminal residues are recognized by a special ubiquitin ligase that is conserved from yeast to humans.

As we have seen, all proteins are initially synthesized bearing methionine (or formylmethionine in bacteria), as their N-terminal residue, which is a stabilizing residue in the N-end rule. Special proteases, called methionine aminopeptidases, will often remove the first methionine of a nascent protein, but they will do so only if the second residue is also stabilizing according to N-end rule. Therefore, it was initially unclear how N-end rule substrates form *in vivo*. However, it is now understood that these substrates are formed by site-specific proteases. For example, a subunit of cohesin, a protein complex that holds sister chromatids together, is cleaved by a highly specific protease during the metaphase–anaphase transition. This cell-cycle-regulated cleavage allows separation of the sister chromatids and leads to the completion of mitosis (see

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 142 of 240

Figure 17–44). The C-terminal fragment of the cleaved subunit bears an N-terminal arginine, a destabilizing residue in the N-end rule. Mutant cells lacking the N-end rule pathway exhibit a greatly increased frequency of chromosome loss, presumably because a failure to degrade this fragment of the cohesin subunit interferes with the formation of new chromatid-associated cohesin complexes in the next cell cycle.

## Abnormally Folded Proteins Can Aggregate to Cause Destructive Human Diseases

Many inherited human diseases (for example, sickle-cell anemia (see p. 1495) and α-1-antitrypsin deficiency, a condition that often leads to liver disease and emphysema) result from mutant proteins that escape the cell's quality controls, fold abnormally, and form aggregates. By absorbing critical macromolecules, these aggregates can severely damage cells and even cause cell death. Often, the inheritance of a single mutant allele of a gene can cause disease, since the normal copy of the gene cannot protect the cell from the destructive properties of the aggregate.

In normal humans, the gradual decline of the cell's protein quality controls can also cause disease by permitting normal proteins to form aggregates (**Figure 6–95**). In some cases, the protein aggregates are released from dead cells and accumulate in the extracellular matrix that surrounds the cells in a tissue, and in extreme cases they can also damage tissues. Because the brain is composed of a highly organized collection of nerve cells, it is especially vulnerable.

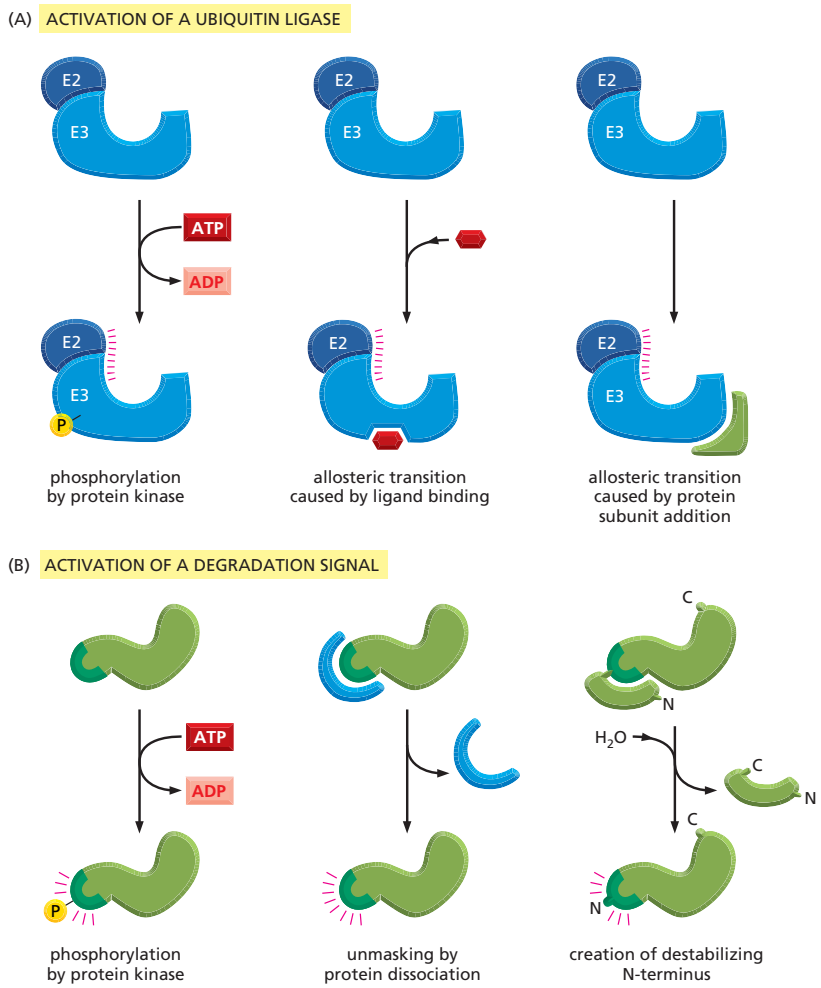



**Figure 6–94 Two general ways of inducing the degradation of a specific protein.** (A) Activation of a specific E3 molecule creates a new ubiquitin ligase. (B) Creation of an exposed degradation signal in the protein to be degraded. This signal binds a ubiquitin ligase, causing the addition of a polyubiquitin chain to a nearby lysine on the target protein. All six pathways shown are known to be used by cells to induce the movement of selected proteins into the proteasome.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 143 of 240

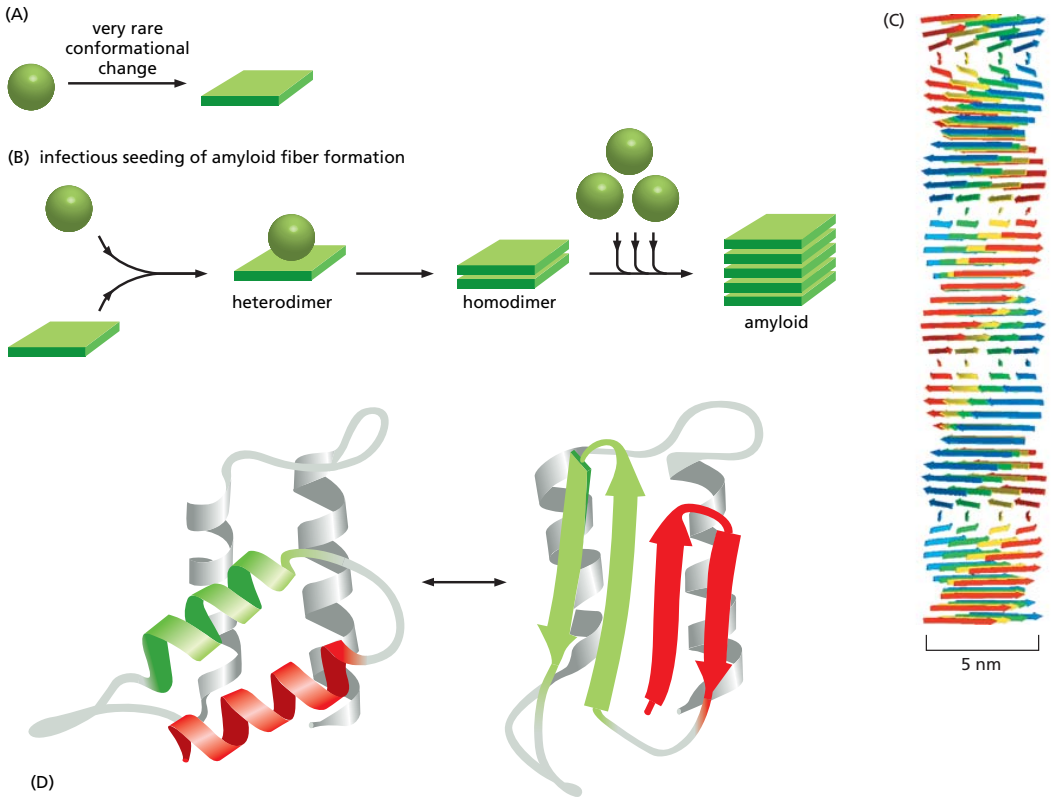



**Figure 6–95 Protein aggregates that cause human disease.** (A) Schematic illustration of the type of conformational change in a protein that produces material for a cross-beta filament. (B) Diagram illustrating the self-infectious nature of the protein aggregation that is central to prion diseases. PrP (prion protein) is highly unusual because the misfolded version of the protein, called PrP*, induces the normal PrP protein it contacts to change its conformation, as shown. Most of the human diseases caused by protein aggregation are caused by the overproduction of a variant protein that is especially prone to aggregation, but the protein aggregate cannot spread from one animal to another. (C) Drawing of a cross-beta filament, a common type of protease-resistant protein aggregate found in many human neurological diseases. Because the hydrogen-bond interactions in a β sheet form between polypeptide backbone atoms (see Figure 3–9), a number of different abnormally folded proteins can produce this structure. (D) One of several possible models for the conversion of PrP to PrP*, showing the likely change of two α helices into four β strands. Although the structure of the normal protein has been determined accurately, the structure of the infectious form is not yet known with certainty because the aggregation has prevented the use of standard structural techniques. (C, courtesy of Louise Serpell, adapted from M. Sunde et al., *J. Mol. Biol.* 273:729–739, 1997. With permission from Academic Press; D, adapted from S.B. Prusiner, *Trends Biochem. Sci.* 21:482–487, 1996. With permission from Elsevier.)

Not surprisingly, therefore, protein aggregates primarily cause neurodegenerative diseases. Prominent among these are Huntington's disease and Alzheimer's disease—the latter causing age-related dementia in more than 20 million people in today's world.

For a particular type of protein aggregate to survive, grow, and damage an organism, it must be highly resistant to proteolysis both inside and outside the cell. Many of the protein aggregates that cause problems form fibrils built from a series of polypeptide chains that are layered one over the other as a continuous stack of β sheets. This so-called *cross-beta filament* (Figure 6–95C), a structure particularly resistant to proteolysis, is observed in many of the neurological disorders caused by protein aggregates, where it produces distinctly staining deposits known as *amyloids*.

One particular variety of these pathologies has attained special notoriety. These are the **prion diseases**. Unlike Huntington's or Alzheimer's, prion diseases can spread from one organism to another, providing that the second organism eats a tissue containing the protein aggregate. A set of diseases—called scrapie in sheep, Creutzfeldt–Jacob disease (CJD) in humans, and bovine spongiform encephalopathy (BSE) in cattle—are caused by a misfolded, aggregated form of a protein called PrP (for prion protein). The PrP is normally located on the outer surface of the plasma membrane, most prominently in neurons. Its normal

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 144 of 240

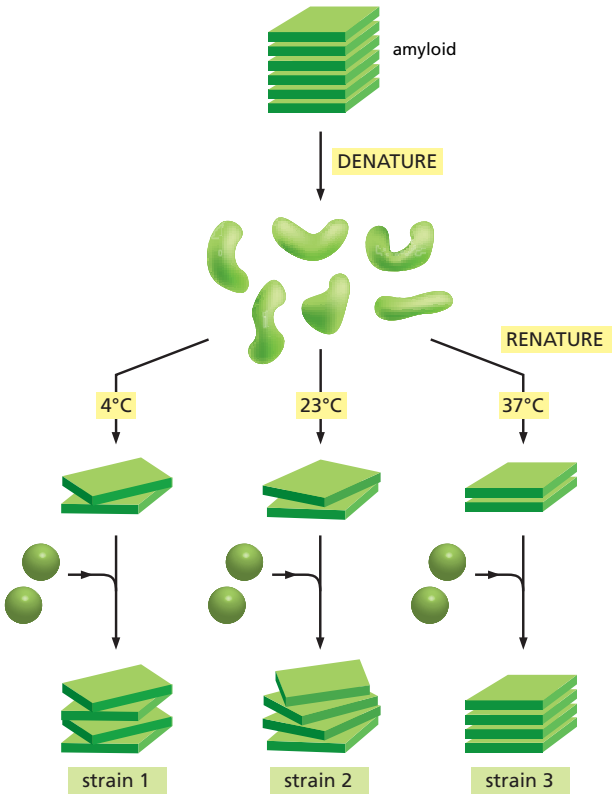



**Figure 6–96 Creation of different prion strains *in vitro*.** In this experiment, amyloid fibers were denatured and the components renatured at different temperatures. This treatment produced three distinctive types of amyloids, each of which could self-propagate when new subunits are added.

function is not known. However, PrP has the unfortunate property of being convertible to a very special abnormal conformation (see Figure 6–95A). This conformation not only forms protease-resistant, cross-beta filaments; it is also "infectious" because it converts normally folded molecules of PrP to the same pathological form. This property creates a positive feedback loop that propagates the abnormal form of PrP, called PrP* (see Figure 6–95B) and thereby allows the pathological conformation to spread rapidly from cell to cell in the brain, eventually causing death in both animals and humans. It can be dangerous to eat the tissues of animals that contain PrP*, as witnessed by the spread of BSE (commonly referred to as "mad cow disease") from cattle to humans in Great Britain. Fortunately, in the absence of PrP*, PrP is extraordinarily difficult to convert to its abnormal form.

Although very few proteins have the potential to misfold into an infectious conformation, another example causes an otherwise mysterious "protein-only inheritance" observed in yeast cells. The ability to study infectious proteins in yeast has clarified another remarkable feature of prions. These protein molecules can form several distinctively different types of aggregates from the same polypeptide chain. Moreover, each type of aggregate can be infectious, forcing normal protein molecules to adopt the same type of abnormal structure. Thus, several different "strains" of infectious particles can arise from the same polypeptide chain (**Figure 6–96**). How a single polypeptide sequence can adopt multiple aggregate forms is not fully understood; it is possible that all prion aggregates resemble cross-beta filaments (see Figure 6–95C) where the structure is held together predominantly with main peptide chain interactions. This would leave the amino acid side chains free to adopt different conformations and, if the structures are self-propagating, the existence of different strains could be explained.

Finally, although prions were discovered because they cause disease, they also appear to have some positive roles in the cell. For example, some species of fungi use prion transformations to establish different types of cells. Although the idea is controversial, it has even been proposed that prions have a role in consolidating memories in complex, multicellular organisms like ourselves.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 145 of 240

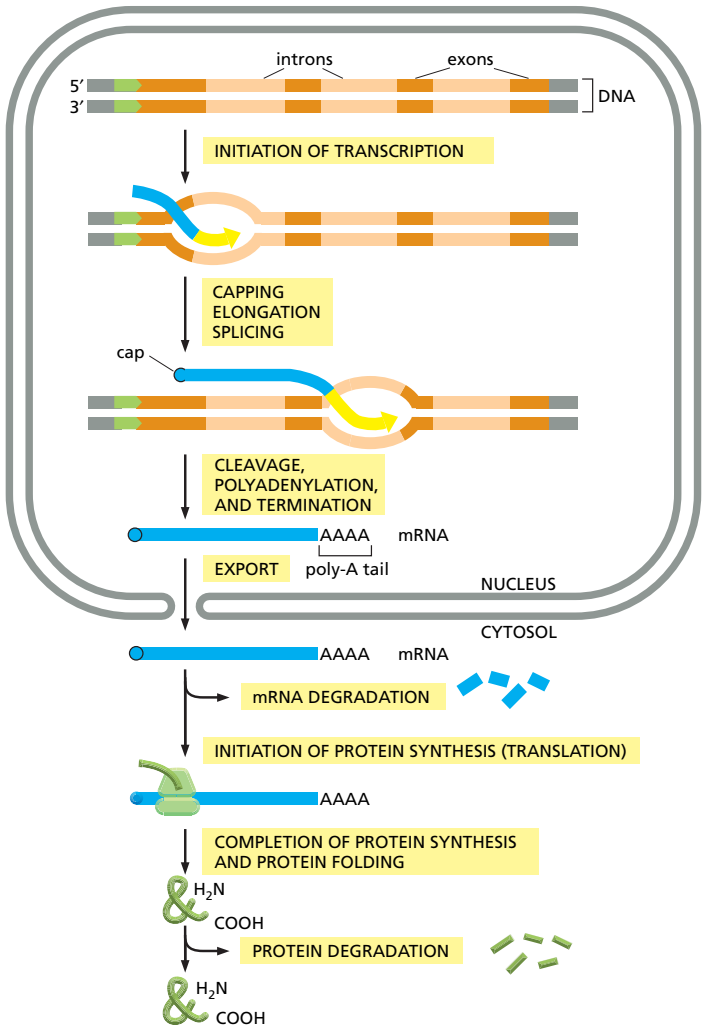



**Figure 6–97 The production of a protein by a eucaryotic cell.** The final level of each protein in a eucaryotic cell depends upon the efficiency of each step depicted.

## There Are Many Steps From DNA to Protein

We have seen so far in this chapter that many different types of chemical reactions are required to produce a properly folded protein from the information contained in a gene (**Figure 6–97**). The final level of a properly folded protein in a cell therefore depends upon the efficiency with which each of the many steps is performed.

In the following chapter, we shall see that cells have the ability to change the levels of their proteins according to their needs. In principle, any or all of the steps in Figure 6–97 could be regulated for each individual protein. As we shall see in Chapter 7, there are examples of regulation at each step from gene to protein. However, the initiation of transcription is the most common point for a cell to regulate the expression of each of its genes. This makes sense, inasmuch as the most efficient way to keep a gene from being expressed is to block the very first step—the transcription of its DNA sequence into an RNA molecule.

## Summary

*The translation of the nucleotide sequence of an mRNA molecule into protein takes place in the cytoplasm on a large ribonucleoprotein assembly called a ribosome. The amino acids used for protein synthesis are first attached to a family of tRNA molecules, each of which recognizes, by complementary base-pair interactions, particular sets of three nucleotides in the mRNA (codons). The sequence of nucleotides in the mRNA is then read from one end to the other in sets of three according to the genetic code.*

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 146 of 240

*To initiate translation, a small ribosomal subunit binds to the mRNA molecule at a start codon (AUG) that is recognized by a unique initiator tRNA molecule. A large ribosomal subunit binds to complete the ribosome and begin protein synthesis. During this phase, aminoacyl-tRNAs—each bearing a specific amino acid—bind sequentially to the appropriate codons in mRNA through complementary base pairing between tRNA anticodons and mRNA codons. Each amino acid is added to the C-terminal end of the growing polypeptide in four sequential steps: aminoacyl-tRNA binding, followed by peptide bond formation, followed by two ribosome translocation steps. Elongation factors use GTP hydrolysis to drive these reactions forward and to improve the accuracy of amino acid selection. The mRNA molecule progresses codon by codon through the ribosome in the 5′-to-3′ direction until it reaches one of three stop codons. A release factor then binds to the ribosome, terminating translation and releasing the completed polypeptide.*

*Eucaryotic and bacterial ribosomes are closely related, despite differences in the number and size of their rRNA and protein components. The rRNA has the dominant role in translation, determining the overall structure of the ribosome, forming the binding sites for the tRNAs, matching the tRNAs to codons in the mRNA, and creating the active site of the peptidyl transferase enzyme that links amino acids together during translation.*

*In the final steps of protein synthesis, two distinct types of molecular chaperones guide the folding of polypeptide chains. These chaperones, known as Hsp60 and Hsp70, recognize exposed hydrophobic patches on proteins and serve to prevent the protein aggregation that would otherwise compete with the folding of newly synthesized proteins into their correct three-dimensional conformations. This protein folding process must also compete with an elaborate quality control mechanism that destroys proteins with abnormally exposed hydrophobic patches. In this case, ubiquitin is covalently added to a misfolded protein by a ubiquitin ligase, and the resulting polyubiquitin chain is recognized by the cap on a proteasome that moves the entire protein to the interior of the proteasome for proteolytic degradation. A closely related proteolytic mechanism, based on special degradation signals recognized by ubiquitin ligases, is used to determine the lifetimes of many normally folded proteins. By this method, selected normal proteins are removed from the cell in response to specific signals.*

## THE RNA WORLD AND THE ORIGINS OF LIFE

We have seen that the expression of hereditary information requires extraordinarily complex machinery and proceeds from DNA to protein through an RNA intermediate. This machinery presents a central paradox: if nucleic acids are required to synthesize proteins and proteins are required, in turn, to synthesize nucleic acids, how did such a system of interdependent components ever arise? One view is that an *RNA world* existed on Earth before modern cells arose (**Figure 6–98**). According to this hypothesis, RNA both stored genetic information and catalyzed the chemical reactions in primitive cells. Only later in evolutionary time did DNA take over as the genetic material and proteins become the major catalyst and structural component of cells. If this idea is correct, then the transition out of the RNA world was never complete; as we have seen in this chapter, RNA still catalyzes several fundamental reactions in modern-day cells, which can be viewed as molecular fossils of an earlier world.

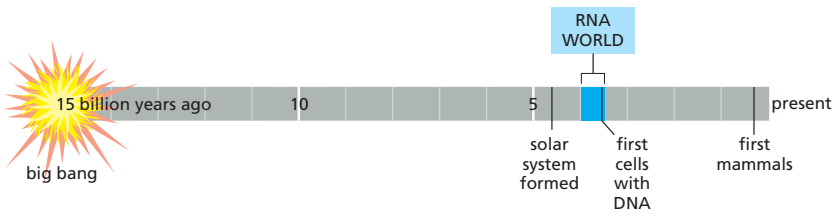



**Figure 6–98** Time line for the universe, suggesting the early existence of an RNA world of living systems.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 147 of 240

In this section we present some of the arguments in support of the RNA world hypothesis. We will see that several of the more surprising features of modern-day cells, such as the ribosome and the pre-mRNA splicing machinery, are most easily explained by viewing them as descendants of a complex network of RNA-mediated interactions that dominated cell metabolism in the RNA world. We also discuss how DNA may have taken over as the genetic material, how the genetic code may have arisen, and how proteins may have eclipsed RNA to perform the bulk of biochemical catalysis in modern-day cells.

## Life Requires Stored Information

It has been proposed that the first "biological" molecules on Earth were formed by metal-based catalysis on the crystalline surfaces of minerals. In principle, an elaborate system of molecular synthesis and breakdown (metabolism) could have existed on these surfaces long before the first cells arose. Although controversial, many scientists believe that an extensive phase of "chemical evolution" took place on the prebiotic Earth, during which small molecules that could catalyze their own synthesis competed with each other for raw materials.

But life requires much more than this. As described in Chapter 1, *heredity* is perhaps the central feature of life. Not only must a cell use raw materials to create a network of catalyzed reactions, it must do so according to an elaborate set of instructions encoded in the hereditary information. The replication of this information ensures that the complex metabolism of cells can accurately reproduce itself. Another crucial feature of life is the genetic variability that results from changes in the hereditary information. This variability, acted upon by selective pressures, is responsible for the great diversity of life on our planet.

Thus, the emergence of life requires a way to store information, a way to duplicate it, a way to change it, and a way to convert the information through catalysis into favorable chemical reactions. But how could such a system begin to be formed? In present-day cells the most versatile catalysts are polypeptides, composed of many different amino acids with chemically diverse side chains and, consequently, able to adopt diverse three-dimensional forms that bristle with reactive chemical groups. Polypeptides also carry information, in the order of their amino acid subunits. But there is no known way in which a polypeptide can reproduce itself by directly specifying the formation of another of precisely the same sequence.

## Polynucleotides Can Both Store Information and Catalyze Chemical Reactions

Polynucleotides have one property that contrasts with those of polypeptides: they can directly guide the formation of copies of their own sequence. This capacity depends on complementary base pairing of nucleotide subunits, which enables one polynucleotide to act as a template for the formation of another. As we have seen in this and the preceding chapter, such complementary templating mechanisms lie at the heart of DNA replication and transcription in modern-day cells.

But the efficient synthesis of polynucleotides by such complementary templating mechanisms requires catalysts to promote the polymerization reaction: without catalysts, polymer formation is slow, error-prone, and inefficient. Today, template-based nucleotide polymerization is rapidly catalyzed by protein enzymes—such as the DNA and RNA polymerases. How could such polymerization be catalyzed before proteins with the appropriate enzymatic specificity existed? The beginnings of an answer to this question came from the discovery in 1982 that RNA molecules themselves can act as catalysts. We have seen in this chapter, for example, that a molecule of RNA catalyzes one of the central reactions in the cell, the covalent joining of amino acids to form proteins. The unique potential of RNA molecules to act both as information carrier and as catalyst forms the basis of the RNA world hypothesis.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 148 of 240

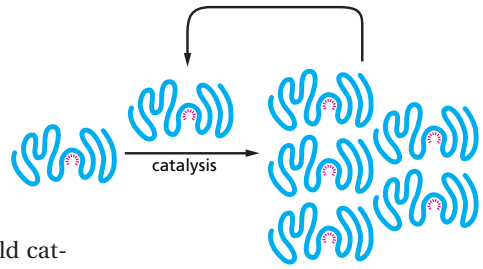



**Figure 6–99 An RNA molecule that can catalyze its own synthesis.** This hypothetical process would require catalysis of the production of both a second RNA strand of complementary nucleotide sequence and the use of this second RNA molecule as a template to form many molecules of RNA with the original sequence. The *red* rays represent the active site of this hypothetical RNA enzyme.

RNA therefore has all the properties required of a molecule that could catalyze a variety of chemical reactions, including those that lead to its own synthesis (**Figure 6–99**). Although self-replicating systems of RNA molecules have not been found in nature, scientists are confident that they can be constructed in the laboratory. While this demonstration would not prove that self-replicating RNA molecules were essential in the origin of life on Earth, it would certainly indicate that such a scenario is possible.

## A Pre-RNA World May Predate the RNA World

Although RNA seems well suited to form the basis for a self-replicating set of biochemical catalysts, it is not clear that RNA was the first kind of molecule to do so. From a purely chemical standpoint, it is difficult to imagine how long RNA molecules could be formed initially by purely nonenzymatic means. For one thing, the precursors of RNA, the ribonucleotides, are difficult to form nonenzymatically. Moreover, the formation of RNA requires that a long series of 3′-to-5′ phosphodiester linkages assemble in the face of a set of competing reactions, including hydrolysis, 2′-to-5′ linkages, and 5′-to-5′ linkages. Given these problems, it has been suggested that the first molecules to possess both catalytic activity and information storage capabilities may have been polymers that resemble RNA but are chemically simpler (**Figure 6–100**). We do not have any remnants of these compounds in present-day cells, nor do such compounds leave fossil records. Nonetheless, the relative simplicity of these "RNA-like polymers" suggests that one of them, rather than RNA itself, may have been the first biopolymer on Earth capable of both information storage and catalytic activity.

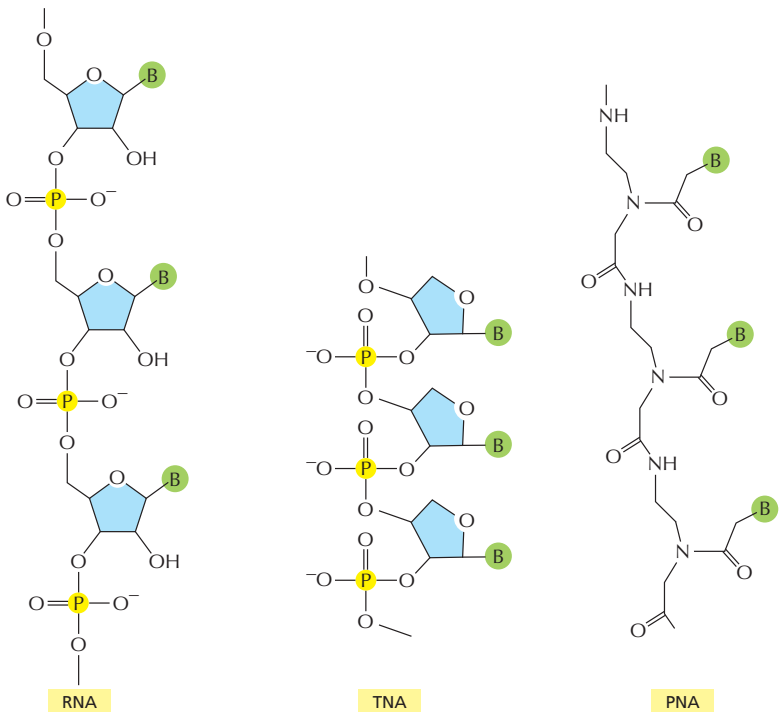



**Figure 6–100 Structures of RNA and two related information-carrying polymers.** In each case, B indicates a purine or pyrimidine base. The polymer TNA (threose nucleic acid) has a 4-carbon sugar unit in contrast to the 5-carbon ribose in RNA. In PNA (peptide nucleic acid), the ribose phosphate backbone of RNA has been replaced by the peptide backbone found in proteins. Like RNA, TNA and PNA can form double helices through complementary base-pairing, and each could therefore in principle serve as a template for its own synthesis.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 149 of 240

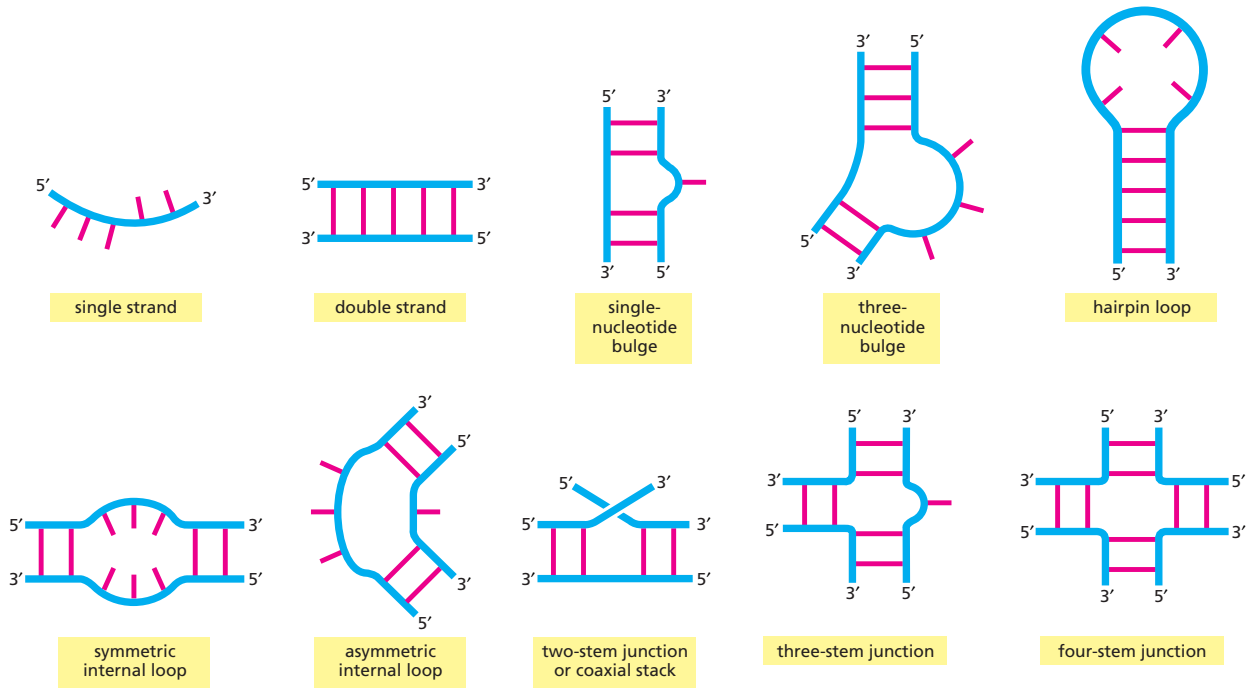



**Figure 6–101 Common elements of RNA secondary structure.** Conventional, complementary base-pairing interactions are indicated by *red* "rungs" in double-helical portions of the RNA.

If the pre-RNA world hypothesis is correct, then a transition to the RNA world must have occurred, presumably through the synthesis of RNA using one of these simpler polymers as both template and catalyst. While the details of the pre-RNA and RNA worlds will likely remain unknown, we know for certain that RNA molecules can catalyze a wide variety of chemical reactions, and we now turn to the properties of RNA that make this possible.

## Single-Stranded RNA Molecules Can Fold into Highly Elaborate Structures

We have seen that complementary base-pairing and other types of hydrogen bonds can occur between nucleotides in the same chain, causing an RNA molecule to fold up in a unique way determined by its nucleotide sequence (see, for example, Figures 6–6, 6–52, and 6–69). Comparisons of many RNA structures have revealed conserved motifs, short structural elements that are used over and over again as parts of larger structures. **Figure 6–101** shows some of these RNA secondary structural motifs, and **Figure 6–102** shows a few common examples of more complex and often longer-range interactions, known as RNA tertiary interactions.

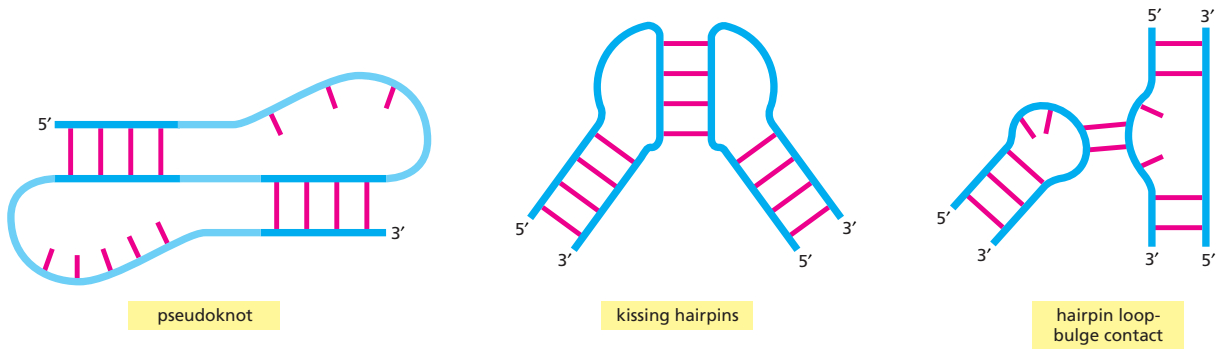



**Figure 6–102 Examples of RNA tertiary interactions.** Some of these interactions can join distant parts of the same RNA molecule or bring two separate RNA molecules together.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 150 of 240

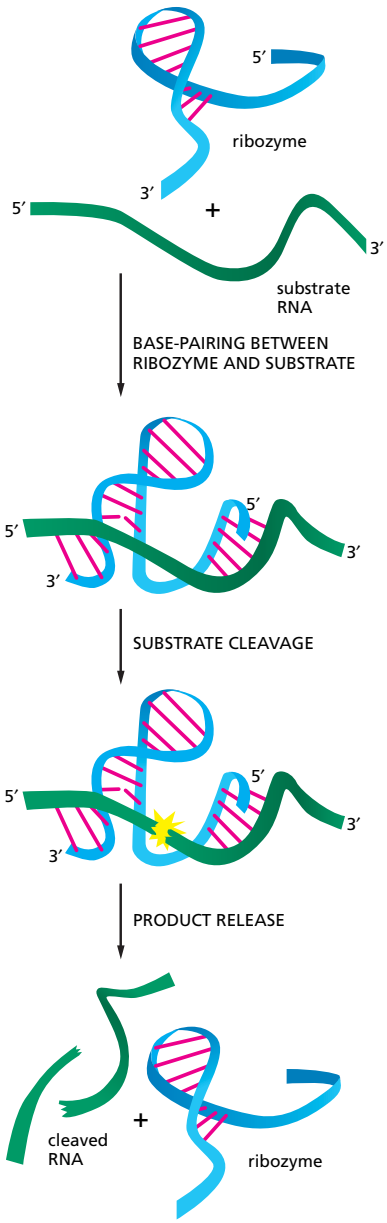



**Figure 6–103 A ribozyme.** This simple RNA molecule catalyzes the cleavage of a second RNA at a specific site. This ribozyme is found embedded in larger RNA genomes—called viroids—which infect plants. The cleavage, which occurs in nature at a distant location on the same RNA molecule that contains the ribozyme, is a step in the replication of the viroid genome. Although not shown in the figure, the reaction requires a Mg molecule positioned at the active site. (Adapted from T.R. Cech and O.C. Uhlenbeck, *Nature* 372:39–40, 1994. With permission from Macmillan Publishers Ltd.)

Protein catalysts require a surface with unique contours and chemical properties on which a given set of substrates can react (discussed in Chapter 3). In exactly the same way, an RNA molecule with an appropriately folded shape can serve as an enzyme (**Figure 6–103**). Like some proteins, many of these ribozymes work by positioning metal ions at their active sites. This feature gives them a wider range of catalytic activities than the limited chemical groups of the polynucleotide chain.

Relatively few catalytic RNAs are known to exist in modern-day cells, however, and much of our inference about the RNA world has come from experiments in which large pools of RNA molecules of random nucleotide sequences are generated in the laboratory. Those rare RNA molecules with a property specified by the experimenter are then selected out and studied (**Figure 6–104**). Such experiments have created RNAs that can catalyze a wide variety of biochemical reactions (**Table 6–5**), with reaction rate enhancements approaching those of proteins. Given these findings, it is not clear why protein catalysts greatly outnumber ribozymes in modern cells. Experiments have shown that RNA molecules may have more difficulty than proteins in binding to flexible, hydrophobic substrates; moreover, the availability of 20 types of amino acids over four types of bases may provide proteins with a greater number of catalytic strategies.

Like proteins, RNAs can undergo conformational changes, either in response to small molecules or to other RNAs. We saw several examples of this in the ribosome and the spliceosome, and we will see others in Chapter 7 when we discuss *riboswitches*. One of the most dramatic RNA conformational changes has been observed with an artificial ribozyme which can exist in two entirely different conformations, each with a different catalytic activity (**Figure 6–105**). Since the discovery of catalysis by RNA, it has become clear that RNA is an enormously versatile molecule, and it is therefore not unreasonable to contemplate the past existence of an RNA world with a very high level of biochemical sophistication.

## Self-Replicating Molecules Undergo Natural Selection

The three-dimensional folded structure of a polynucleotide affects its stability, its actions on other molecules, and its ability to replicate. Therefore, certain polynucleotides will be especially successful in any self-replicating mixture. Because errors inevitably occur in any copying process, new variant sequences of these polynucleotides will be generated over time.

Certain catalytic activities would have had a cardinal importance in the early evolution of life. Consider in particular an RNA molecule that helps to catalyze the process of templated polymerization, taking any given RNA molecule as a template (**Figure 6–106**). Such a molecule, by acting on copies of itself, can replicate. At the same time, it can promote the replication of other types of RNA molecules in its neighborhood (**Figure 6–107**). If some of these neighboring RNAs have catalytic actions that help the survival of RNA in other ways (catalyzing ribonucleotide production, for example), a set of different types of RNA molecules, each specialized for a different activity, could evolve into a cooperative system that replicates with unusually great efficiency.

But for any of these cooperative systems to evolve, they must be present together in a compartment. For example, a set of mutually beneficial RNAs (such

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 151 of 240

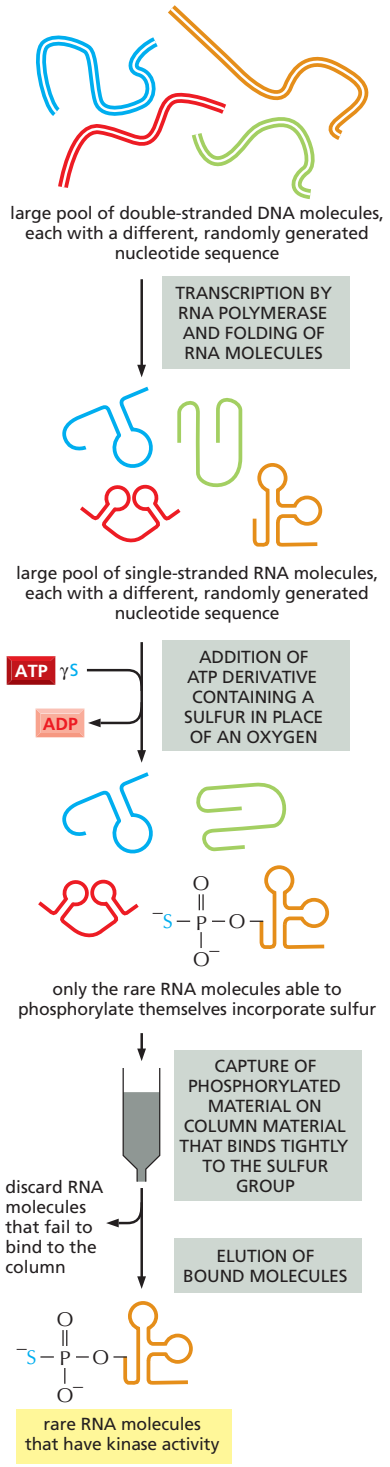



**Figure 6–104** ***In vitro*** **selection of a synthetic ribozyme.** Beginning with a large pool of nucleic acid molecules synthesized in the laboratory, those rare RNA molecules that possess a specified catalytic activity can be isolated and studied. Although a specific example (that of an autophosphorylating ribozyme) is shown, variations of this procedure have been used to generate many of the ribozymes listed in Table 6–5. During the autophosphorylation step, the RNA molecules are kept sufficiently dilute to prevent the "cross"-phosphorylation of additional RNA molecules. In reality, several repetitions of this procedure are necessary to select the very rare RNA molecules with this catalytic activity. Thus, the material initially eluted from the column is converted back into DNA, amplified many fold (using reverse transcriptase and PCR as explained in Chapter 8), transcribed back into RNA, and subjected to repeated rounds of selection. (Adapted from J.R. Lorsch and J.W. Szostak, *Nature* 371:31–36, 1994. With permission from Macmillan Publishers Ltd.)

as those of Figure 6–107) could replicate themselves only if all the RNAs remained in the neighborhood of the RNA that is specialized for templated polymerization. Moreover, compartmentalization would bar parasitic RNA molecules from entering the system. Selection of a set of RNA molecules according to the quality of the self-replicating systems they generated could not therefore occur efficiently until some form of compartment evolved to contain them.

An early, crude form of compartmentalization may have been simple adsorption on surfaces or particles. The need for more sophisticated types of containment is easily fulfilled by a class of small molecules that has the simple physicochemical property of being *amphiphilic*, that is, consisting of one part that is hydrophobic (water insoluble) and another part that is hydrophilic (water soluble). When such molecules are placed in water, they aggregate, arranging their hydrophobic portions as much in contact with one another as possible and their hydrophilic portions in contact with the water. Amphiphilic molecules of appropriate shape aggregate spontaneously to form *bilayers*, creating small closed vesicles whose aqueous contents are isolated from the external medium (**Figure 6–108**). The phenomenon can be demonstrated in a test tube by simply mixing phospholipids and water together: under appropriate conditions, small vesicles will form. All present-day cells are surrounded by a *plasma membrane* consisting of amphiphilic molecules—mainly phospholipids—in this configuration; we discuss these molecules in detail in Chapter 10.

The spontaneous assembly of a set of amphiphilic molecules, enclosing a self-replicating mixture of RNAs (or pre-RNAs) and other molecules (**Figure**

**Table 6–5 Some Biochemical Reactions That Can Be Catalyzed by Ribozymes**

| ACTIVITY | RIBOZYMES |
|---|---|
| Peptide bond formation in protein synthesis | ribosomal RNA |
| RNA cleavage, RNA ligation | self-splicing RNAs; RNase P; also *in vitro* selected RNA |
| DNA cleavage | self-splicing RNAs |
| RNA splicing | self-splicing RNAs, perhaps RNAs of the spliceosome |
| RNA polymerizaton | *in vitro* selected RNA |
| RNA and DNA phosphorylation | *in vitro* selected RNA |
| RNA aminoacylation | *in vitro* selected RNA |
| RNA alkylation | *in vitro* selected RNA |
| Amide bond formation | *in vitro* selected RNA |
| Glycosidic bond formation | *in vitro* selected RNA |
| Oxidation/reduction reactions | *in vitro* selected RNA |
| Carbon–carbon bond formation | *in vitro* selected RNA |
| Phosphoamide bond formation | *in vitro* selected RNA |
| Disulfide exchange | *in vitro* selected RNA |

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 152 of 240

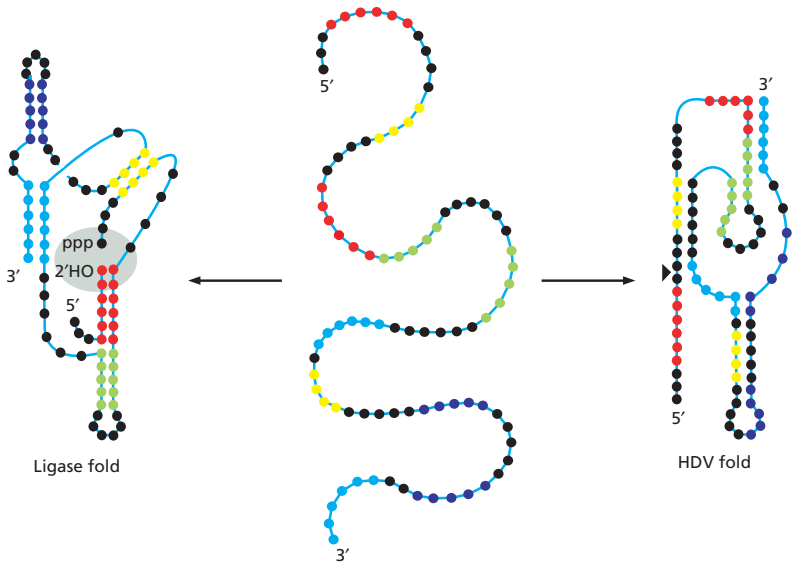



**Figure 6–105 An RNA molecule that folds into two different ribozymes.** This 88-nucleotide RNA, created in the laboratory, can fold into a ribozyme that carries out a self-ligation reaction *(left)* or a self-cleavage reaction *(right)*. The ligation reaction forms a 2′,5′ phosphodiester linkage with the release of pyrophosphate. This reaction seals the gap *(gray* shading), which was experimentally introduced into the RNA molecule. In the reaction carried out by the HDV fold, the RNA is cleaved at this same position, indicated by the *arrowhead*. This cleavage resembles that used in the life cycle of HDV, a hepatitis B satellite virus, hence the name of the fold. Each nucleotide is represented by a *colored dot*, with the colors used simply to clarify the two different folding patterns. The folded structures illustrate the secondary structures of the two ribozymes with regions of base-pairing indicated by close oppositions of the *colored dots*. Note that the two ribozyme folds have no secondary structure in common. (Adapted from E.A. Schultes and D.P. Bartel, *Science* 289:448–452, 2000. With permission from AAAS.)

**6–109**), presumably formed the first membrane-bounded cells. Although it is not clear at what point in the evolution of biological catalysts this might have occurred, once RNA molecules were sealed within a closed membrane they could begin to evolve in earnest as carriers of genetic instructions: new variants could be selected not merely on the basis of their own structure, but also according to their effect on the other molecules in the same compartment. The nucleotide sequences of the RNA molecules could now be expressed in the character of a unitary living cell.

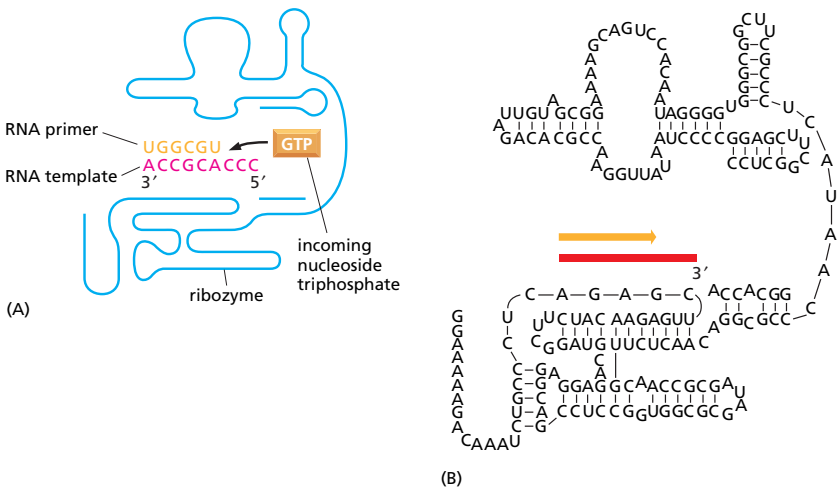



**Figure 6–106 A ribozyme created in the laboratory that can catalyze templated synthesis of RNA from nucleoside triphosphates.** (A) Schematic diagram of the ribozyme showing one step of the templated polymerization reaction it catalyzes. (B) Nucleotide sequence of the ribozyme with base pairings indicated. Although relatively inefficient (it can only synthesize short lengths of RNA), this ribozyme adds the correct base, as specified by the template, over 95% of the time. (From W.K. Johnston et al., *Science* 292:1319–1325, 2001. With permission from AAAS.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 153 of 240

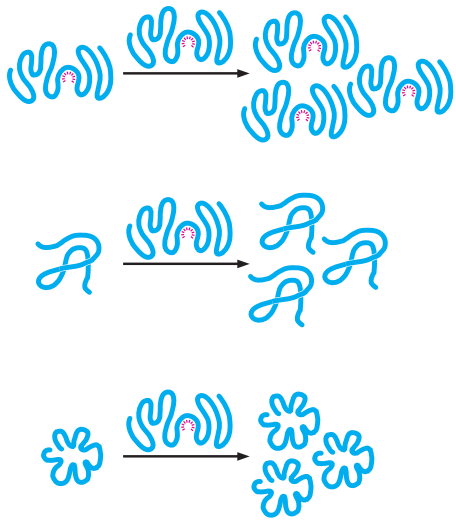


**Figure 6–107 A family of mutually supportive RNA molecules.** One molecule is a ribozyme that replicates itself as well as the other RNA molecules. The other molecules would catalyze secondary tasks needed for the survival of the cooperative system, for example, by synthesizing ribonucleotides for RNA synthesis or phospholipids for compartmentalization.

## How Did Protein Synthesis Evolve?

The molecular processes underlying protein synthesis in present-day cells seem inextricably complex. Although we understand most of them, they do not make conceptual sense in the way that DNA transcription, DNA repair, and DNA replication do. It is especially difficult to imagine how protein synthesis evolved because it is now performed by a complex interlocking system of protein and RNA molecules; obviously the proteins could not have existed until an early version of the translation apparatus was already in place. The RNA world hypothesis is especially appealing because the use of RNA in both information and catalysis seems both economic and conceptually simple. As attractive as this idea is for envisioning early life, it does not explain how the modern-day system of protein synthesis arose. Although we can only speculate on the origins of modern protein synthesis and the genetic code, several experimental observations have provided plausible scenarios.

In modern cells, some short peptides (such as antibiotics) are synthesized without the ribosome; peptide synthetase enzymes assemble these peptides, with their proper sequence of amino acids, without mRNAs to guide their synthesis. It is plausible that this non-coded, primitive version of protein synthesis first developed during the RNA world where it would have been catalyzed by RNA molecules. This idea presents no conceptual difficulties because, as we have seen, rRNA catalyzes peptide bond formation in present-day cells. We also know that ribozymes created in the laboratory can perform specific aminoacylation reactions; that is, they can match specific amino acids to specific tRNAs. It is therefore possible that tRNA-like adapters, each matched to a specific amino acid, could have arisen in the RNA world, marking the beginnings of a genetic code.

In principle, other RNAs (the precursors to mRNAs) could have served as crude templates to direct the nonrandom polymerization of a few different amino acids. Any RNA that helped guide the synthesis of a useful polypeptide would have a great advantage in the evolutionary struggle for survival. We can envision a relatively nonspecific peptidyl transferase ribozyme, which, over time, grew larger and acquired the ability to position charged tRNAs accurately on RNA templates—leading eventually to the modern ribosome. Once protein

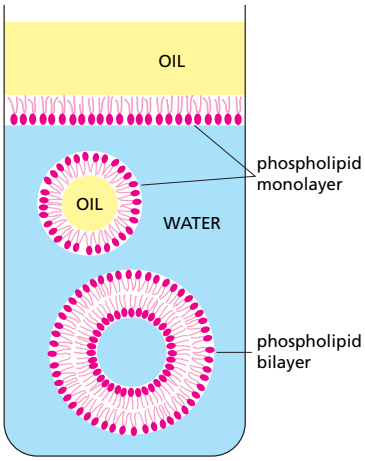



**Figure 6–108 Formation of membrane by phospholipids.** Because these molecules have hydrophilic heads and lipophilic tails, they align themselves at an oil/water interface with their heads in the water and their tails in the oil. In the water they associate to form closed bilayer vesicles in which the lipophilic tails are in contact with one another and the hydrophilic heads are exposed to the water.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 154 of 240

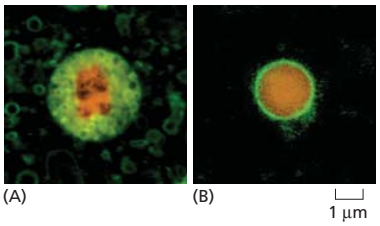



**Figure 6–109 Encapsulation of RNA by simple amphiphilic molecules.** For these experiments, the clay mineral montmorillonite was used to bring together RNA and fatty acids. (A) A montmorillonite particle, coated by RNA (*red*) has become trapped inside a fatty acid vesicle (*green*). (B) RNA *(red)* in solution has been encapsulated by fatty acids *(green)*. These experiments show that montmorillonite can greatly accelerate the spontaneous generation of vesicles from amphiphilic molecules and trap RNA inside them. It has been hypothesized that conceptually similar actions may have led to the first primitive cells on Earth. (From M.M. Hanczyc et al., *Science* 302:618–622, 2003. With permission from AAAS.)

synthesis evolved, the transition to a protein-dominated world could proceed, with proteins eventually taking over the majority of catalytic and structural tasks because of their greater versatility, with 20 rather than 4 different subunits. Although the scenarios just discussed are highly speculative, the known properties of RNA molecules are consistent with these ideas.

## All Present-Day Cells Use DNA as Their Hereditary Material

If the evolutionary speculations embodied in the RNA world hypothesis are correct, early cells would have differed fundamentally from the cells we know today in having their hereditary information stored in RNA rather than in DNA (**Figure 6–110**). Evidence that RNA arose before DNA in evolution can be found in the chemical differences between them. Ribose, like glucose and other simple carbohydrates, can be formed from formaldehyde (HCHO), a simple chemical which is readily produced in laboratory experiments that attempt to simulate conditions on the primitive Earth. The sugar deoxyribose is harder to make, and in present-day cells it is produced from ribose in a reaction catalyzed by a protein enzyme, suggesting that ribose predates deoxyribose in cells. Presumably, DNA appeared on the scene later, but then proved more suitable than RNA as a permanent repository of genetic information. In particular, the deoxyribose in its sugar-phosphate backbone makes chains of DNA chemically more stable than chains of RNA, so that much greater lengths of DNA can be maintained without breakage.

The other differences between RNA and DNA—the double-helical structure of DNA and the use of thymine rather than uracil—further enhance DNA stability by making the many unavoidable accidents that occur to the molecule much easier to repair, as discussed in detail in Chapter 5 (see pp. 296–297 and 300–301).

## Summary

*From our knowledge of present-day organisms and the molecules they contain, it seems likely that the development of the directly autocatalytic mechanisms fundamental to living systems began with the evolution of families of molecules that could catalyze their own replication. With time, a family of cooperating RNA catalysts probably developed the ability to direct the synthesis of polypeptides. DNA is likely to have been a late addition: as the accumulation of additional protein catalysts allowed more efficient and complex cells to evolve, the DNA double helix replaced RNA as a more stable molecule for storing the increased amounts of genetic information required by such cells.*

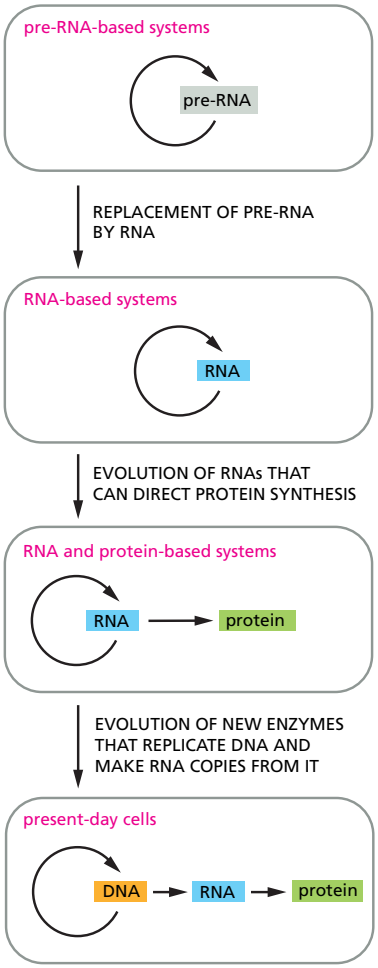



**Figure 6–110 The hypothesis that RNA preceded DNA and proteins in evolution.** In the earliest cells, pre-RNA molecules would have had combined genetic, structural, and catalytic functions and RNA would have gradually taken over these functions. In present-day cells, DNA is the repository of genetic information, and proteins perform the vast majority of catalytic functions in cells. RNA primarily functions today as a go-between in protein synthesis, although it remains a catalyst for a small number of crucial reactions.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 155 of 240

# PROBLEMS

## Which statements are true? Explain why or why not.

**6–1** The consequences of errors in transcription are less than those of errors in RNA replication.

**6–2** Since introns are largely genetic "junk," they do not have to be removed precisely from the primary transcript during RNA splicing.

**6–3** Wobble pairing occurs between the first position in the codon and the third position in the anticodon.

**6–4** Protein enzymes are thought to greatly outnumber ribozymes in modern cells because they catalyze a much greater variety of reactions at much faster rates than ribozymes.

## Discuss the following problems.

**6–5** In which direction along the template must the RNA polymerase in **Figure Q6–1** be moving to have generated the supercoiled structures that are shown? Would you expect supercoils to be generated if the RNA polymerase were free to rotate about the axis of the DNA as it progressed along the template?

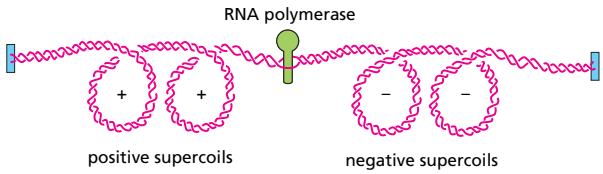



**Figure Q6–1** Supercoils around a moving RNA polymerase (Problem 6–5).

**6–6** Phosphates are attached to the CTD (C-terminal domain) of RNA polymerasse II during transcription. What are the various roles of RNA polymerase II CTD phosphorylation?

**6–7** The human α-tropomyosin gene is alternatively spliced to produce several forms of α-tropomyosin mRNA in various cell types (**Figure Q6–2**). For all forms of the mRNA, the encoded protein sequence is the same for exons 1 and

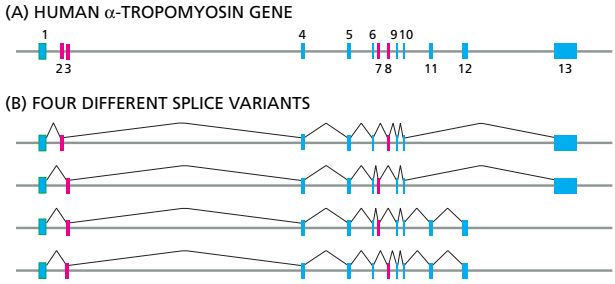



**Figure Q6–2** Alternatively spliced mRNAs from the human α-tropomyosin gene (Problem 6–7). (A) Exons in the human α-tropomyosin gene. The locations and relative sizes of exons are shown by the *blue* and *red rectangles*. (B) Splicing patterns for four α-tropomyosin mRNAs. Splicing is indicated by *lines* connecting the exons that are included in the mRNA.

10. Exons 2 and 3 are alternative exons used in different mRNAs, as are exons 7 and 8. Which of the following statements about exons 2 and 3 is the most accurate? Is that statement also the most accurate one for exons 7 and 8? Explain your answers.

**A.** Exons 2 and 3 must have the same number of nucleotides.

**B.** Exons 2 and 3 must each contain an integral number of codons (that is, the number of nucleotides divided by 3 must be an integer).

**C.** Exons 2 and 3 must each contain a number of nucleotides that when divided by 3 leaves the same remainder (that is, 0, 1, or 2).

**6–8** After treating cells with a chemical mutagen, you isolate two mutants. One carries alanine and the other carries methionine at a site in the protein that normally contains valine (**Figure Q6–3**). After treating these two mutants again with the mutagen, you isolate mutants from each that now carry threonine at the site of the original valine (Figure Q6–3). Assuming that all mutations involve single nucleotide changes, deduce the codons that are used for valine, methionine, threonine, and alanine at the affected site. Would you expect to be able to isolate valine-to-threonine mutants in one step?

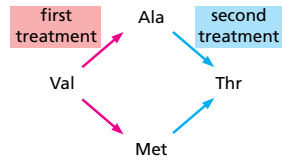



**Figure Q6–3** Two rounds of mutagenesis and the altered amino acids at a single position in a protein (Problem 6–8).

**6–9** The elongation factor EF-Tu introduces two short delays between codon–anticodon base-pairing and formation of the peptide bond. These delays increase the accuracy of protein synthesis. Describe these delays and explain how they improve the fidelity of translation.

**6–10** Both Hsp60-like and Hsp70 molecular chaperones share an affinity for exposed hydrophobic patches on proteins, using them as indicators of incomplete folding. Why do you suppose hydrophobic patches serve as critical signals for the folding status of a protein?

**6–11** Most proteins require molecular chaperones to assist in their correct folding. How do you suppose the chaperones themselves manage to fold correctly?

**6–12** What is so special about RNA that makes it such an attractive evolutionary precursor to DNA and protein? What is it about DNA that makes it a better material than RNA for storage of genetic information?

**6–13** If an RNA molecule could form a hairpin with a symmetric internal loop, as shown in **Figure Q6–4**, could the complement of this RNA form a similar structure? If so, would there be any regions of the two structures that are identical? Which ones?

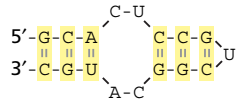



**Figure Q6–4** An RNA hairpin with a symmetric internal loop (Problem 6–13).

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 156 of 240

# REFERENCES


### General

Berg JM, Tymoczko JL & Stryer L (2006) Biochemistry, 6th ed. New York: WH Freeman.
Brown TA (2002) Genomes 2, 2nd ed. New York: Wiley-Liss.
Gesteland RF, Cech TR & Atkins JF (eds) (2006) The RNA World, 3rd ed. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
Hartwell L, Hood L, Goldberg ML et al (2006) Genetics: from Genes to Genomes, 3rd ed. Boston: McGraw Hill.
Lodish H, Berk A, Kaiser C et al (2007) Molecular Cell Biology, 6th ed. New York: WH Freeman.
Stent GS (1971) Molecular Genetics: An Introductory Narrative. San Francisco: WH Freeman.
The Genetic Code (1966) *Cold Spring Harb. Symp Quant Biol* 31.
The Ribosome (2001) *Cold Spring Harb Symp Quant Biol* 66.
Watson JD, Baker TA, Bell SP et al (2003) Molecular Biology of the Gene, 5th ed. Menlo Park, CA: Benjamin Cummings.

### From DNA to RNA

Bentley DL (2005) Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors. *Curr Opin Cell Biol* 17:251–256.
Berget SM, Moore C & Sharp PA (1977) Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci USA* 74:3171–3175.
Black DL (2003). Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* 72:291–336.
Brenner S, Jacob F & Meselson M (1961) An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature* 190:576–581.
Cate JH, Gooding AR, Podell E et al. (1996) Crystal structure of a group I ribozyme domain: principles of RNA packing. *Science* 273:1678–1685.
Chow LT, Gelinas RE, Broker TR et al (1977) An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. *Cell* 12:1–8.
Cramer P (2002) Multisubunit RNA polymerases. *Curr Opin Struct Biol* 12:89–97.
Daneholt B (1997) A look at messenger RNP moving through the nuclear pore. *Cell* 88:585–588.
Dreyfuss G, Kim VN, & Kataoka N (2002) Messenger-RNA-binding proteins and the messages they carry. *Nature Rev Mol Cell Biol* 3:195–205.
Houseley J, LaCava J & Tollervey D (2006) RNA-quality control by the exosome. *Nature Rev Mol Cell Biol* 7:529–539.
Izquierdo JM, & Valcárcel J (2006) A simple principle to explain the evolution of pre-mRNA splicing. *Genes Dev* 20:1679–1684.
Kornberg RD (2005) Mediator and the mechanism of transcriptional activation. *Trends Biochem Sci* 30:235–239.
Malik S & Roeder RG (2005) Dynamic regulation of pol II transcription by the mammalian Mediator complex. *Trends Biochem Sci* 30:256–263.
Matsui T, Segall J, Weil PA & Roeder RG (1980) Multiple factors required for accurate initiation of transcription by purified RNA polymerase II. *J Biol Chem* 255:11992–11996.
Patel AA & Steitz JA (2003) Splicing double: insights from the second spliceosome. *Nature Rev Mol Cell Biol* 4|:960–970.
Phatnani HP & Greenleaf AL (2006) Phosphorylation and functions of the RNA polymerase II CTD. *Genes Dev* 20:2922–2936.
Query CC & Konarska MM (2006) Splicing fidelity revisited. *Nature Struct Mol Biol* 13:472–474.
Ruskin B, Krainer AR, Maniatis T et al (1984) Excision of an intact intron as a novel lariat structure during pre-mRNA splicing *in vitro. Cell* 38:317–331.
Spector DL (2003) The dynamics of chromosome organization and gene regulation. *Annu Rev Biochem* 72:573–608.
Staley JP & Guthrie C (1998) Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell* 92:315–326.
Thomas MC & Chiang CM (2006) The general transcription machinery and general cofactors. *Critical Rev Biochem Mol Biol* 41:105–178.
Wang D, Bushnell DA, Westover KD et al (2006) Structural basis of transcription: role of the trigger loop in substrate specificity and catalysis. *Cell* 127:941–954.

### From RNA to Protein

Allen GS & Frank J (2007) Structural insights on the translation initiation complex: ghosts of a universal initiation complex. *Mol Microbiol* 63:941–950.
Anfinsen CB (1973) Principles that govern the folding of protein chains. *Science* 181:223–230.
Brunelle JL, Youngman EM, Sharma D et al (2006) The interaction between C75 of tRNA and the A loop of the ribosome stimulates peptidyl transferase activity. *RNA* 12:33–39.
Chien P, Weissman JS, & DePace AH (2004). Emerging principles of conformation-based prion inheritance. *Annu Rev Biochem* 73:617–656.
Crick FHC (1966) The genetic code: III. *Sci Am* 215:55–62.
Hershko A, Ciechanover A & Varshavsky A (2000) The ubiquitin system. *Nature Med* 6:1073–1081.
Ibba M & Soll D (2000) Aminoacyl-tRNA synthesis. *Annu Rev Biochem* 69:617–650.
Kozak M (1992) Regulation of translation in eukaryotic systems. *Annu Rev Cell Biol* 8:197–225.
Kuzmiak HA, & Maquat LE (2006) Applying nonsense-mediated mRNA decay research to the clinic: progress and challenges. *Trends Mol Med* 12:306–316.
Moore PB & Steitz TA (2005) The ribosome revealed. *Trends Biochem Sci* 30:281–283.
Noller HF (2005) RNA structure: reading the ribosome. *Science* 309:1508–1514.
Ogle JM, Carter AP & Ramakrishnan V (2003) Insights into the decoding mechanism from recent ribosome structures. *Trends Biochem Sci* 28:259–266.
Prusiner SB (1998) Nobel lecture. Prions. *Proc Natl Acad Sci USA* 95:13363–13383.
Rehwinkel J, Raes J & Izaurralde E (2006) Nonsense-mediated mRNA decay: Target genes and functional diversification of effectors. *Trends Biochem Sci* 31:639–646.
Sauer RT, Bolon DN, Burton BM et al (2004) Sculpting the proteome with AAA(+) proteases and disassembly machines. *Cell* 119:9–18.
Shorter J & Lindquist S (2005) Prions as adaptive conduits of memory and inheritance. *Nature Rev Genet* 6:435–450.
Varshavsky A (2005) Regulated protein degradation. *Trends in Biochem Sci* 30:283–286.
Voges D, Zwickl P & Baumeister W (1999) The 26S proteasome: a molecular machine designed for controlled proteolysis. *Annu Rev Biochem* 68:1015–1068.
Weissmann C (2005) Birth of a prion: spontaneous generation revisited. *Cell* 122:165–168.
Young JC, Agashe VR, Siegers K et al (2004) Pathways of chaperone-mediated protein folding in the cytosol. *Nature Rev Mol Cell Biol* 5:781–791.

### The RNA World and the Origins of Life

Joyce GF (1992) Directed molecular evolution. *Sci Am* 267:90–97.
Orgel L (2000) Origin of life. A simpler nucleic acid. *Science* 290:1306–1307.
Kruger K, Grabowski P, Zaug P et al (1982) Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening seuence of Tetrahymena. *Cell* 31:147–157.
Silverman SK (2003) Rube Goldberg goes (ribo)nuclear? Molecular switches and sensors made from RNA. *RNA* 9:377–383.
Szostak JW, Bartel DP & Luisi PL (2001) Synthesizing life. *Nature* 409:387–390.


LCY Biotechnology Holding, Inc.
Ex. 1008
Page 157 of 240

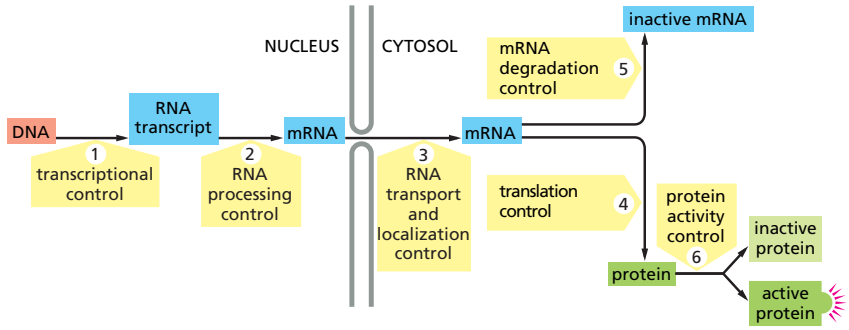



**Figure 7–5 Six steps at which eucaryotic gene expression can be controlled.** Controls that operate at steps 1 through 5 are discussed in this chapter. Step 6, the regulation of protein activity, occurs largely through covalent post-translational modifications including phosphorylation, acetylation, and ubiquitylation (see Table 3–3, p. 186) and is discussed in many chapters throughout the book.

production of glucose from amino acids and other small molecules; the set of proteins whose production is induced includes enzymes such as tyrosine aminotransferase, which helps to convert tyrosine to glucose. When the hormone is no longer present, the production of these proteins drops to its normal level.

Other cell types respond to glucocorticoids differently. Fat cells, for example, reduce the production of tyrosine aminotransferase, while some other cell types do not respond to glucocorticoids at all. These examples illustrate a general feature of cell specialization: different cell types often respond differently to the same extracellular signal. Underlying such adjustments that occur in response to extracellular signals, there are features of the gene expression pattern that do not change and give each cell type its permanently distinctive character.

## Gene Expression Can Be Regulated at Many of the Steps in the Pathway from DNA to RNA to Protein

If differences among the various cell types of an organism depend on the particular genes that the cells express, at what level is the control of gene expression exercised? As we saw in the previous chapter, there are many steps in the pathway leading from DNA to protein. We now know that all of them can in principle be regulated. Thus a cell can control the proteins it makes by (1) controlling when and how often a given gene is transcribed (**transcriptional control**), (2) controlling the splicing and processing of RNA transcripts (**RNA processing control**), (3) selecting which completed mRNAs are exported from the nucleus to the cytosol and determining where in the cytosol they are localized (**RNA transport and localization control**), (4) selecting which mRNAs in the cytoplasm are translated by ribosomes (**translational control**), (5) selectively destabilizing certain mRNA molecules in the cytoplasm (**mRNA degradation control**), or (6) selectively activating, inactivating, degrading, or locating specific protein molecules after they have been made (**protein activity control**) (**Figure 7–5**).

For most genes transcriptional controls are paramount. This makes sense because, of all the possible control points illustrated in Figure 7–5, only transcriptional control ensures that the cell will not synthesize superfluous intermediates. In the following sections we discuss the DNA and protein components that perform this function by regulating the initiation of gene transcription. We shall return at the end of the chapter to the many additional ways of regulating gene expression.

## Summary

*The genome of a cell contains in its DNA sequence the information to make many thousands of different protein and RNA molecules. A cell typically expresses only a fraction of its genes, and the different types of cells in multicellular organisms arise because different sets of genes are expressed. Moreover, cells can change the pattern of genes they express in response to changes in their environment, such as signals from other cells. Although all of the steps involved in expressing a gene can in principle be regulated, for most genes the initiation of RNA transcription is the most important point of control.*

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 158 of 240

# Manipulating Proteins, DNA, and RNA

# 8

Progress in science is often driven by advances in technology. The entire field of cell biology, for example, came into being when optical craftsmen learned to grind small lenses of sufficiently high quality to observe cells and their substructures. Innovations in lens grinding, rather than any conceptual or philosophical advance, allowed Hooke and van Leeuwenhoek to discover a previously unseen cellular world, where tiny creatures tumble and twirl in a small droplet of water (**Figure 8–1**).

The 21st century promises to be a particularly exciting time for biology. New methods for analyzing proteins, DNA, and RNA are fueling an information explosion and allowing scientists to study cells and their macromolecules in previously unimagined ways. We now have access to the sequences of many billions of nucleotides, providing the complete molecular blueprints for hundreds of organisms—from microbes and mustard weeds to worms, flies, mice, dogs, chimpanzees, and humans. And powerful new techniques are helping us to decipher that information, allowing us not only to compile huge, detailed catalogs of genes and proteins but also to begin to unravel how these components work together to form functional cells and organisms. The long-range goal is nothing short of obtaining a complete understanding of what takes place inside a cell as it responds to its environment and interacts with its neighbors. We want to know which genes are switched on, which mRNA transcripts are present, and which proteins are active—where they are located, with what other proteins and other molecules they associate, and to which pathways or networks they belong. We also want to understand how the cell successfully manages this staggering number of variables and how it chooses among an almost unlimited number of possibilities in performing its diverse biological tasks. Such information will permit us to begin to build a framework for delineating, and eventually predicting, how genes and proteins operate to lay the foundations for life.

In this chapter, we present some of the principal methods used to study the molecular components of cells, particularly proteins, DNA, and RNA. We consider how to separate cells of different types from tissues, how to grow cells outside the body, and how to disrupt cells and isolate their organelles and constituent macromolecules in pure form. We also present the latest techniques used to determine protein structure, function, and interactions, and we discuss the breakthroughs in DNA technology that continue to revolutionize our understanding of cell function.

The techniques and methods described in this chapter have made possible the discoveries that are presented throughout this book, and they are currently being used by tens of thousands of scientists each day.

## In This Chapter


ISOLATING CELLS AND GROWING THEM IN CULTURE 501

PURIFYING PROTEINS 510

ANALYZING PROTEINS 517

ANALYZING AND MANIPULATING DNA 532

STUDYING GENE EXPRESSION AND FUNCTION 553

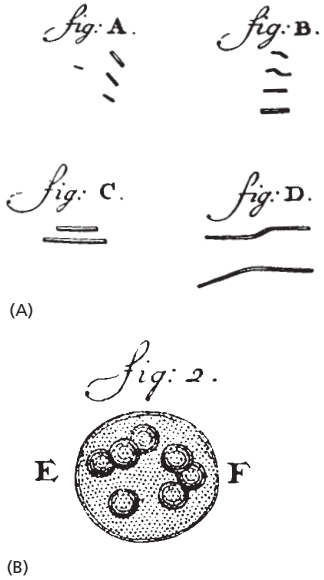




**Figure 8–1 Microscopic life.** A sample of "diverse animalcules" seen by van Leeuwenhoek using his simple microscope. (A) Bacteria seen in material he excavated from between his teeth. Those in fig. B he described as "swimming first forward and then backwards" (1692). (B) The eucaryotic green alga *Volvox* (1700). (Courtesy of the John Innes Foundation.)

## ISOLATING CELLS AND GROWING THEM IN CULTURE

Although the organelles and large molecules in a cell can be visualized with microscopes, understanding how these components function requires a detailed biochemical analysis. Most biochemical procedures require that large



LCY Biotechnology Holding, Inc.
Ex. 1008
Page 159 of 240

numbers of cells be physically disrupted to gain access to their components. If the sample is a piece of tissue, composed of different types of cells, heterogeneous cell populations will be mixed together. To obtain as much information as possible about the cells in a tissue, biologists have developed ways of dissociating cells from tissues and separating them according to type. These manipulations result in a relatively homogeneous population of cells that can then be analyzed—either directly or after their number has been greatly increased by allowing the cells to proliferate in culture.

## Cells Can Be Isolated from Intact Tissues

Intact tissues provide the most realistic source of material, as they represent the actual cells found within the body. The first step in isolating individual cells is to disrupt the extracellular matrix and cell–cell junctions that hold the cells together. For this purpose, a tissue sample is typically treated with proteolytic enzymes (such as trypsin and collagenase) to digest proteins in the extracellular matrix and with agents (such as ethylenediaminetetraacetic acid, or EDTA) that bind, or chelate, the $Ca^{2+}$ on which cell–cell adhesion depends. The tissue can then be teased apart into single cells by gentle agitation.

For some biochemical preparations, the protein of interest can be obtained in sufficient quantity without having to separate the tissue or organ into cell types. Examples include the preparation of histones from calf thymus, actin from rabbit muscle, or tubulin from cow brain. In other cases, obtaining the desired protein requires enrichment for a specific cell type of interest. Several approaches are used to separate the different cell types from a mixed cell suspension. The most general cell-separation technique uses an antibody coupled to a fluorescent dye to label specific cells. An antibody is chosen that specifically binds to the surface of only one cell type in the tissue. The labeled cells can then be separated from the unlabeled ones in an electronic *fluorescence-activated cell sorter*. In this remarkable machine, individual cells traveling single file in a fine stream pass through a laser beam, and the fluorescence of each cell is rapidly measured. A vibrating nozzle generates tiny droplets, most containing either one cell or no cells. The droplets containing a single cell are automatically given a positive or a negative charge at the moment of formation, depending on whether the cell they contain is fluorescent; they are then deflected by a strong electric field into an appropriate container. Occasional clumps of cells, detected by their increased light scattering, are left uncharged and are discarded into a waste container. Such machines can accurately select 1 fluorescent cell from a pool of 1000 unlabeled cells and sort several thousand cells each second (**Figure 8–2**).

Selected cells can also be obtained by carefully dissecting them from thin tissue slices that have been prepared for microscopic examination (discussed in Chapter 9). In one approach, a tissue section is coated with a thin plastic film and a region containing the cells of interest is irradiated with a focused pulse from an infrared laser. This light pulse melts a small circle of the film, binding the cells underneath. These captured cells are then removed for further analysis. The technique, called *laser capture microdissection*, can be used to separate and analyze cells from different areas of a tumor, allowing their properties or molecular composition to be compared with neighboring normal cells. A related method uses a laser beam to directly cut out a group of cells and catapult them into an appropriate container for future analysis (**Figure 8–3**).

A uniform population of cells obtained by any of these or other separation methods can be used directly for biochemical analysis. After breaking open the cells by mechanical disruption, detergents, and other methods, cytoplasm or individual organelles can be extracted and then specific molecules purified.

## Cells Can Be Grown in Culture

Although molecules can be extracted from whole tissues, this is often not the most convenient source of material, requiring, for example, early-morning trips

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 160 of 240

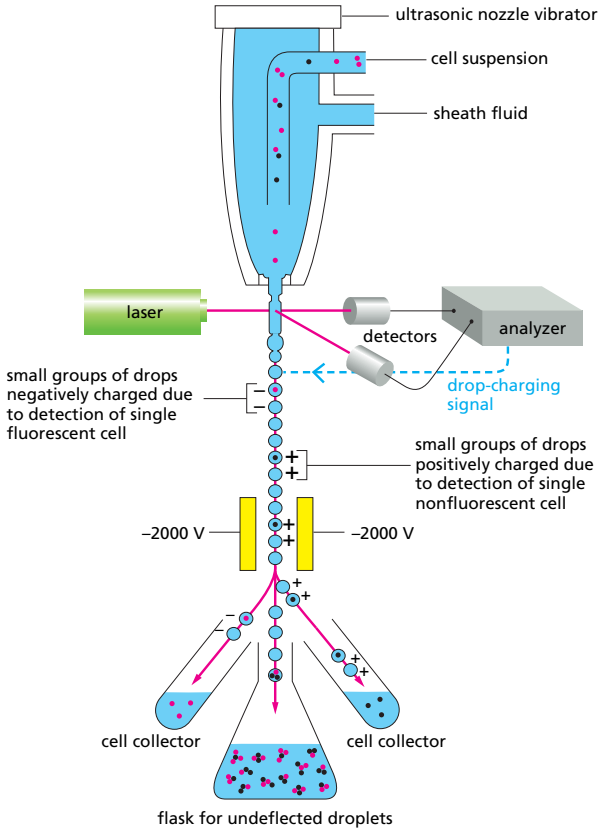



**Figure 8–2 A fluorescence-activated cell sorter.** A cell passing through the laser beam is monitored for fluorescence. Droplets containing single cells are given a negative or positive charge, depending on whether the cell is fluorescent or not. The droplets are then deflected by an electric field into collection tubes according to their charge. Note that the cell concentration must be adjusted so that most droplets contain no cells and flow to a waste container together with any cell clumps.

to a slaughterhouse. The problem is not only a question of convenience. The livestock commonly used as organ sources are not amenable to genetic manipulation. Moreover, the complexity of intact tissues and organs is an inherent disadvantage when trying to purify particular molecules. Cells grown in culture provide a more homogeneous population of cells from which to extract material, and they are also much more convenient to work with in the laboratory. Given appropriate surroundings, most plant and animal cells can live, multiply, and even express differentiated properties in a tissue-culture dish. The cells can be watched continuously under the microscope or analyzed biochemically, and the effects of adding or removing specific molecules, such as hormones or growth factors, can be systematically explored. In addition, by mixing two cell types, the interactions between one cell type and another can be studied.

Experiments performed on cultured cells are sometimes said to be carried out *in vitro* (literally, "in glass") to contrast them with experiments using intact organisms, which are said to be carried out *in vivo* (literally, "in the living organism"). These terms can be confusing, however, because they are often used in a very different sense by biochemists. In the biochemistry lab, *in vitro* refers to reactions carried out in a test tube in the absence of living cells, whereas *in vivo*

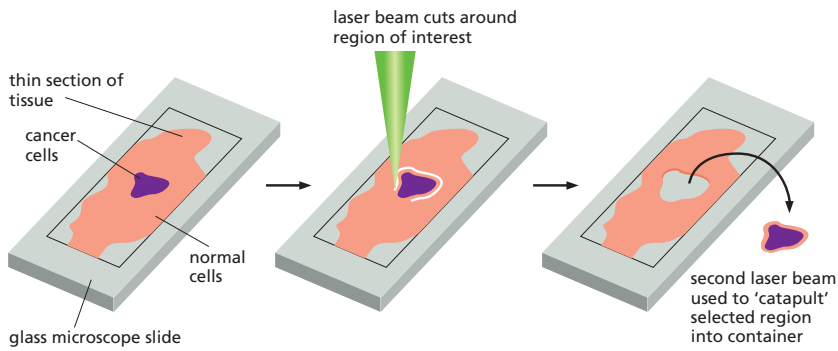
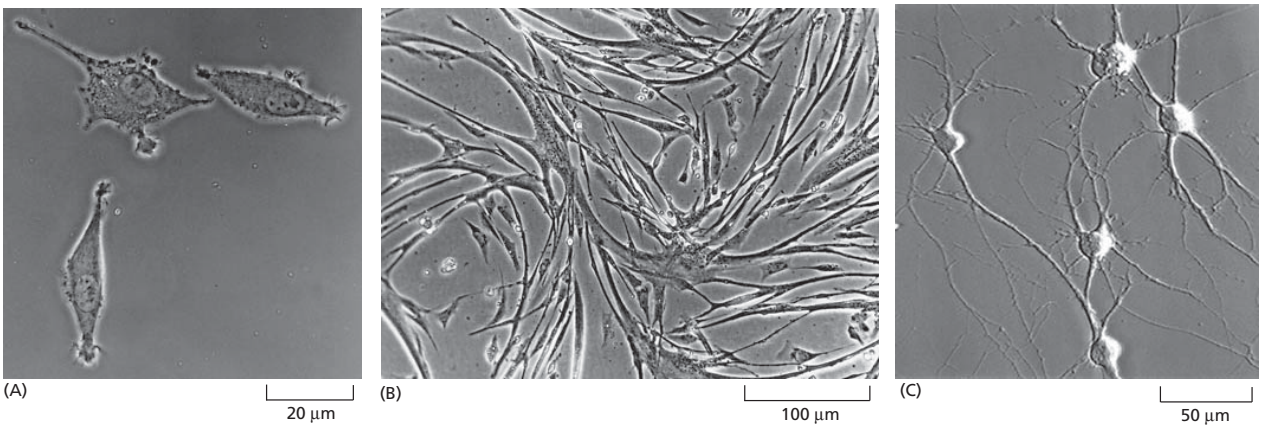



**Figure 8–3 Microdissection techniques to select cells from tissue slices.** This method uses a laser beam to excise a region of interest and eject it into a container, and it permits the isolation of even a single cell from a tissue sample.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 161 of 240



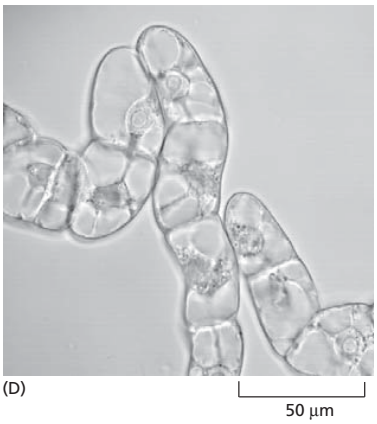




**Figure 8–4 Light micrographs of cells in culture.** (A) Mouse fibroblasts. (B) Chick myoblasts fusing to form multinucleate muscle cells. (C) Purified rat retinal ganglion nerve cells. (D) Tobacco cells in liquid culture. (A, courtesy of Daniel Zicha; B, courtesy of Rosalind Zalin; C, from A. Meyer-Franke et al., *Neuron* 15:805–819, 1995. With permission from Elsevier; D, courtesy of Gethin Roberts.)

refers to any reaction taking place inside a living cell, even if that cell is growing in culture.

Tissue culture began in 1907 with an experiment designed to settle a controversy in neurobiology. The hypothesis under examination was known as the neuronal doctrine, which states that each nerve fiber is the outgrowth of a single nerve cell and not the product of the fusion of many cells. To test this contention, small pieces of spinal cord were placed on clotted tissue fluid in a warm, moist chamber and observed at regular intervals under the microscope. After a day or so, individual nerve cells could be seen extending long, thin filaments (axons) into the clot. Thus, the neuronal doctrine received strong support, and the foundation was laid for the cell-culture revolution.

These original experiments on nerve fibers used cultures of small tissue fragments called explants. Today, cultures are more commonly made from suspensions of cells dissociated from tissues using the methods described earlier. Unlike bacteria, most tissue cells are not adapted to living suspended in fluid and require a solid surface on which to grow and divide. For cell cultures, this support is usually provided by the surface of a plastic tissue-culture dish. Cells vary in their requirements, however, and many do not proliferate or differentiate unless the culture dish is coated with materials that cells like to adhere to, such as polylysine or extracellular matrix components.

Cultures prepared directly from the tissues of an organism are called *primary cultures*. These can be made with or without an initial fractionation step to separate different cell types. In most cases, cells in primary cultures can be removed from the culture dish and recultured repeatedly in so-called secondary cultures; in this way, they can be repeatedly subcultured *(passaged)* for weeks or months. Such cells often display many of the differentiated properties appropriate to their origin (**Figure 8–4**): fibroblasts continue to secrete collagen; cells derived from embryonic skeletal muscle fuse to form muscle fibers that contract spontaneously in the culture dish; nerve cells extend axons that are electrically excitable and make synapses with other nerve cells; and epithelial cells form extensive sheets with many of the properties of an intact epithelium. Because these properties are maintained in culture, they are accessible to study in ways that are often not possible in intact tissues.

Cell culture is not limited to animal cells. When a piece of plant tissue is cultured in a sterile medium containing nutrients and appropriate growth regulators, many of the cells are stimulated to proliferate indefinitely in a disorganized manner, producing a mass of relatively undifferentiated cells called a *callus*. If the nutrients and growth regulators are carefully manipulated, one can induce the formation of a shoot and then root apical meristems within the callus, and,

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 162 of 240

in many species, regenerate a whole new plant. Similar to animal cells, callus cultures can be mechanically dissociated into single cells, which will grow and divide as a suspension culture (see Figure 8–4D).

## Eucaryotic Cell Lines Are a Widely Used Source of Homogeneous Cells

The cell cultures obtained by disrupting tissues tend to suffer from a problem—eventually the cells die. Most vertebrate cells stop dividing after a finite number of cell divisions in culture, a process called *replicative cell senescence* (discussed in Chapter 17). Normal human fibroblasts, for example, typically divide only 25–40 times in culture before they stop. In these cells, the limited proliferation capacity reflects a progressive shortening and uncapping of the cell's telomeres, the repetitive DNA sequences and associated proteins that cap the ends of each chomosome (discussed in Chapter 5). Human somatic cells in the body have turned off production of the enzyme, called *telomerase,* that normally maintains the telomeres, which is why their telomeres shorten with each cell division. Human fibroblasts can often be coaxed to proliferate indefinitely by providing them with the gene that encodes the catalytic subunit of telomerase; in this case, they can be propagated as an "immortalized" cell line.

Some human cells, however, cannot be immortalized by this trick. Although their telomeres remain long, they still stop dividing after a limited number of divisions because the culture conditions eventually activate cell-cycle *checkpoint mechanisms* (discussed in Chapter 17) that arrest the cell cycle—a process sometimes called "culture shock." In order to immortalize these cells, one has to do more than introduce telomerase. One must also inactivate the checkpoint mechanisms. This can be done by introducing certain cancer-promoting oncogenes, such as those derived from tumor viruses (discussed in Chapter 20). Unlike human cells, most rodent cells do not turn off production of telomerase and therefore their telomeres do not shorten with each cell division. Therefore, if culture shock can be avoided, some rodent cell types will divide indefinitely in culture. In addition, rodent cells often undergo genetic changes in culture that inactivate their checkpoint mechanisms, thereby spontaneously producing immortalized cell lines.

Cell lines can often be most easily generated from cancer cells, but these cultures differ from those prepared from normal cells in several ways, and are referred to as *transformed cell lines.* Transformed cell lines often grow without attaching to a surface, for example, and they can proliferate to a much higher density in a culture dish. Similar properties can be induced experimentally in normal cells by transforming them with a tumor-inducing virus or chemical. The resulting transformed cell lines can usually cause tumors if injected into a susceptible animal (although it is usually only a small subpopulation, called cancer stem cells, that can do so—discussed in Chapter 20).

Both transformed and nontransformed cell lines are extremely useful in cell research as sources of very large numbers of cells of a uniform type, especially since they can be stored in liquid nitrogen at –196°C for an indefinite period and retain their viability when thawed. It is important to keep in mind, however, that the cells in both types of cell lines nearly always differ in important ways from their normal progenitors in the tissues from which they were derived.

Some widely used cell lines are listed in **Table 8–1**. Different lines have different advantages; for example, the PtK epithelial cell lines derived from the rat kangaroo, unlike many other cell lines which round up during mitosis, remain flat during mitosis, allowing the mitotic apparatus to be readily observed in action.

## Embryonic Stem Cells Could Revolutionize Medicine

Among the most promising cell lines to be developed—from a medical point of view—are embryonic stem (ES) cells. These remarkable cells, first harvested from the inner cell mass of the early mouse embryo, can proliferate indefinitely

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 163 of 240

**Table 8–1 Some Commonly Used Cell Lines**

| CELL LINE* | CELL TYPE AND ORIGIN |
|---|---|
| 3T3 | fibroblast (mouse) |
| BHK21 | fibroblast (Syrian hamster) |
| MDCK | epithelial cell (dog) |
| HeLa | epithelial cell (human) |
| PtK1 | epithelial cell (rat kangaroo) |
| L6 | myoblast (rat) |
| PC12 | chromaffin cell (rat) |
| SP2 | plasma cell (mouse) |
| COS | kidney (monkey) |
| 293 | kidney (human); transformed with adenovirus |
| CHO | ovary (Chinese hamster) |
| DT40 | lymphoma cell for efficient targeted recombination (chick) |
| R1 | embryonic stem cell (mouse) |
| E14.1 | embryonic stem cell (mouse) |
| H1, H9 | embryonic stem cell (human) |
| S2 | macrophage-like cell *(Drosophila)* |
| BY2 | undifferentiated meristematic cell (tobacco) |

*Many of these cell lines were derived from tumors. All of them are capable of indefinite replication in culture and express at least some of the special characteristics of their cell's of origin.

in culture and yet retain an unrestricted developmental potential. If the cells from the culture dish are put back into an early embryonic environment, they can give rise to all the cell types in the body, including germ cells (**Figure 8–5**). Their descendants in the embryo are able to integrate perfectly into whatever site they come to occupy, adopting the character and behavior that normal cells would show at that site.

Cells with properties similar to those of mouse ES cells can now be derived from early human embryos, creating a potentially inexhaustible supply of cells that might be used to replace and repair damaged mature human tissue. Experiments in mice suggest that it may be possible, in the future, to use ES cells to produce specialized cells for therapy—to replace the skeletal muscle fibers that degenerate in victims of muscular dystrophy, the nerve cells that die in patients with Parkinson's disease, the insulin-secreting cells that are destroyed in type I

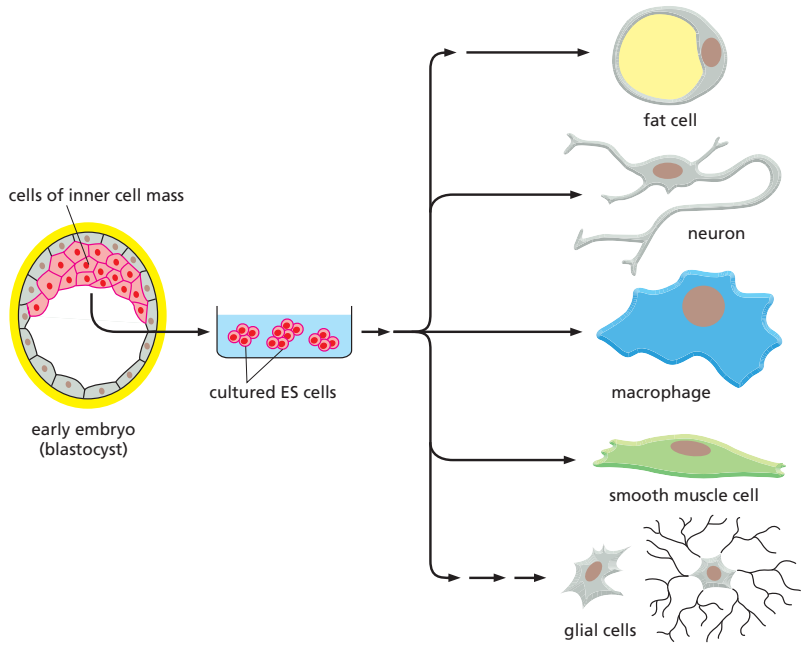



**Figure 8–5 Embryonic stem (ES) cells derived from an embryo.** These cultured cells can give rise to all of the cell types of the body. ES cells are harvested from the inner cell mass of an early embryo and can be maintained indefinitely as stem cells (discussed in Chapter 23) in culture. If they are put back into an embryo, they will integrate perfectly and differentiate to suit whatever environment they find themselves. The cells can also be kept in culture as an immortal cell line; they can then be supplied with different hormones or growth factors to encourage them to differentiate into specific cell types. (Based on E. Fuchs and J.A. Segré, *Cell* 100:143–155, 2000. With permission from Elsevier.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 164 of 240

diabetics, and the cardiac muscle cells that die during a heart attack. Perhaps one day it may even become possible to grow entire organs from ES cells by a recapitulation of embryonic development. It is important not to transplant ES cells by themselves into adults, as they can produce tumors called teratomas.

There is another major problem associated with the use of ES-cell-derived cells for tissue repair. If the transplanted cells differ genetically from the cells of the patient into whom they are grafted, the patient's immune system will reject and destroy those cells. This problem can be avoided, of course, if the cells used for repair are derived from the patient's own body. As discussed in Chapter 23, many adult tissues contain stem cells dedicated to continuous production of just one or a few specialized cell types, and a great deal of stem-cell research aims to manipulate the behavior of these adult stem cells for use in tissue repair.

ES cell technology, however, in theory at least, also offers another way around the problem of immune rejection, using a strategy known as "therapeutic cloning," as we explain next.

## Somatic Cell Nuclear Transplantation May Provide a Way to Generate Personalized Stem Cells

The term "cloning" has been used in confusing ways as a shorthand term for several quite distinct types of procedures. It is important to understand the distinctions, particularly in the context of public debates about the ethics of stem cell research.

As biologists define the term, a *clone* is simply a set of individuals that are genetically identical because they have descended from a single ancestor. The simplest type of cloning is the cloning of cells. Thus, one can take a single epidermal stem cell from the skin and let it grow and divide in culture to obtain a large clone of genetically identical epidermal cells, which can, for example, be used to help reconstruct the skin of a badly burned patient. This kind of cloning is no more than an extension by artificial means of the processes of cell proliferation and repair that occur in a normal human body.

The cloning of entire multicellular animals, called *reproductive cloning,* is a very different enterprise, involving a far more radical departure from the ordinary course of nature. Normally, each individual animal has both a mother and a father, and is not genetically identical to either of them. In reproductive cloning, the need for two parents and sexual union is bypassed. For mammals, this difficult feat has been achieved in sheep and some other domestic animals by *somatic cell nuclear transplantation*. The procedure begins with an unfertilized egg cell. The nucleus of this haploid cell is sucked out and replaced by a nucleus from a regular diploid somatic cell. The diploid donor cell is typically taken from a tissue of an adult individual. The hybrid cell, consisting of a diploid donor nucleus in a host egg cytoplasm, is allowed to develop for a short while in culture. In a small proportion of cases, this procedure can give rise to an early embryo, which is then put into the uterus of a foster mother (**Figure 8–6**). If the experimenter is lucky, development continues like that of a normal embryo,

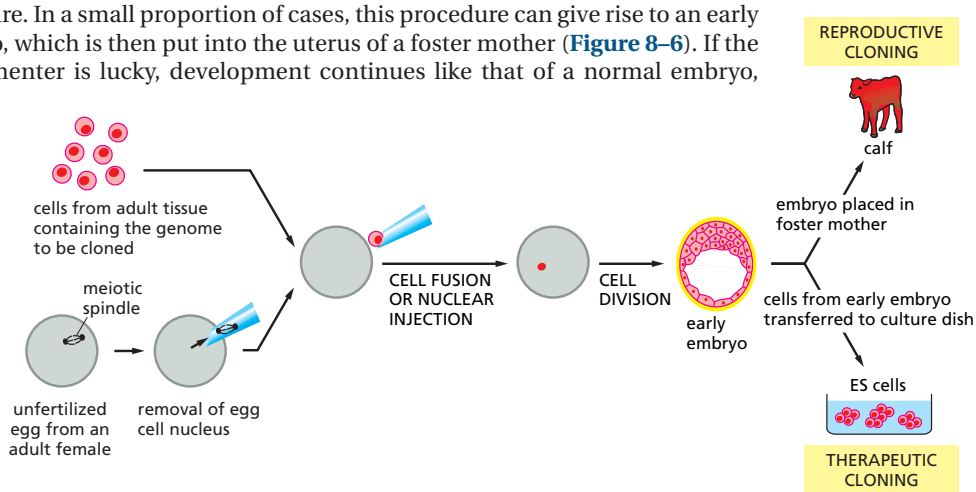



**Figure 8–6 Reproductive and therapeutic cloning.** Cells from adult tissue can be used for reproductive cloning or for generating personalized ES cells (so-called therapeutic cloning).

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 165 of 240

giving rise, eventually, to a whole new animal. An individual produced in this way, by reproductive cloning, should be genetically identical to the adult individual that donated the diploid cell (except for the small amount of genetic information in mitochondria, which is inherited solely from the egg cytoplasm).

Therapeutic cloning, which is very different from reproductive cloning, employs the technique of somatic cell nuclear transplantation to produce personalized ES cells (see Figure 8–6). In this case, the very early embryo generated by nuclear transplantation, consisting of about 200 cells, is not transferred to the uterus of a foster mother. Instead, it is used as a source from which ES cells are derived in culture, with the aim of generating various cell types that can be used for tissue repair. Cells obtained by this route are genetically nearly identical to the donor of the original nucleus, so they can be grafted back into the donor, without fear of immunological rejection.

Somatic cell nuclear transfer has an additional potential benefit—for studying inherited human diseases. ES cells that have received a somatic nucleus from an individual with an inherited disorder can be used to directly study the way in which the disease develops as the ES cells are induced to differentiate into distinct cell types. "Disease-specific" ES cells and their differentiated progeny can also be used to study the progression of such diseases and to test and develop new drugs to treat the disorders. These strategies are still in their infancy, and some countries outlaw certain aspects of the research. It remains to be seen whether human ES cells can be produced by nuclear transfer and whether human ES cells will fulfill the great hopes that medical scientists have for them.

## Hybridoma Cell Lines Are Factories That Produce Monoclonal Antibodies

As we see throughout this book, antibodies are particularly useful tools for cell biology. Their great specificity allows precise visualization of selected proteins among the many thousands that each cell typically produces. Antibodies are often produced by inoculating animals with the protein of interest and subsequently isolating the antibodies specific to that protein from the serum of the animal. However, only limited quantities of antibodies can be obtained from a single inoculated animal, and the antibodies produced will be a heterogeneous mixture of antibodies that recognize a variety of different antigenic sites on a macromolecule that differs from animal to animal. Moreover, antibodies specific for the antigen will constitute only a fraction of the antibodies found in the serum. An alternative technology, which allows the production of an infinite quantity of identical antibodies and greatly increases the specificity and convenience of antibody-based methods, is the production of monoclonal antibodies by hybridoma cell lines.

This technology, developed in 1975, has revolutionized the production of antibodies for use as tools in cell biology, as well as for the diagnosis and treatment of certain diseases, including rheumatoid arthritis and cancer. The procedure requires hybrid cell technology (**Figure 8–7**), and it involves propagating a clone of cells from a single antibody-secreting B lymphocyte to obtain a homogeneous preparation of antibodies in large quantities. B lymphocytes normally have a limited life-span in culture, but individual antibody-producing B lymphocytes from an immunized mouse or rat, when fused with cells derived from a transformed B lymphocyte cell line, can give rise to hybrids that have both the ability to make a particular antibody and the ability to multiply indefinitely in culture. These **hybridomas** are propagated as individual clones, each of which provides a permanent and stable source of a single type of **monoclonal antibody** (**Figure 8–8**). Each type of monoclonal antibody recognizes a single type of antigenic site—for example, a particular cluster of five or six amino acid side chains on the surface of a protein. Their uniform specificity makes monoclonal antibodies much more useful than conventional antisera for most purposes.

An important advantage of the hybridoma technique is that monoclonal antibodies can be made against molecules that constitute only a minor component of a complex mixture. In an ordinary antiserum made against such a mix-

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 166 of 240

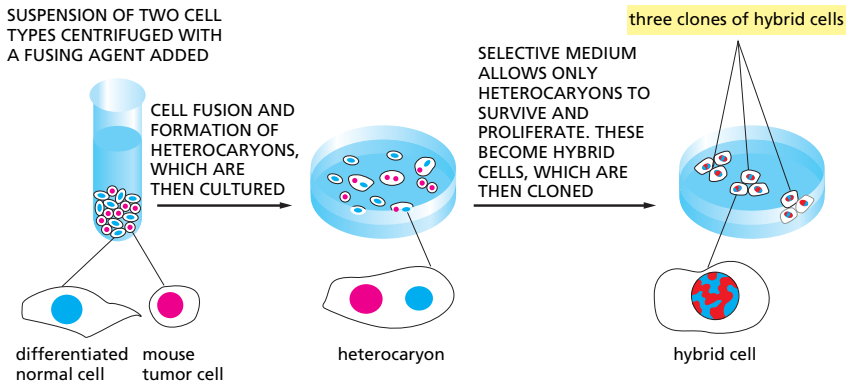



**Figure 8–7 The production of hybrid cells.** It is possible to fuse one cell with another to form a *heterocaryon,* a combined cell with two separate nuclei. Typically, a suspension of cells is treated with certain inactivated viruses or with polyethylene glycol, each of which alters the plasma membranes of cells in a way that induces them to fuse. Eventually, a heterocaryon proceeds to mitosis and produces a hybrid cell in which the two separate nuclear envelopes have been disassembled, allowing all the chromosomes to be brought together in a single large nucleus. Such hybrid cells can give rise to immortal hybrid cell lines. If one of the parent cells was from a tumor cell line, the hybrid cell is called a hybridoma.

ture, the proportion of antibody molecules that recognize the minor component would be too small to be useful. But if the B lymphocytes that produce the various components of this antiserum are made into hybridomas, it becomes possible to screen individual hybridoma clones from the large mixture to select one that produces the desired type of monoclonal antibody and to propagate the selected hybridoma indefinitely so as to produce that antibody in unlimited quantities. In principle, therefore, a monoclonal antibody can be made against any protein in a biological sample. Once an antibody has been made, it can be used to localize the protein in cells and tissues, to follow its movement, and to purify the protein to study its structure and function.

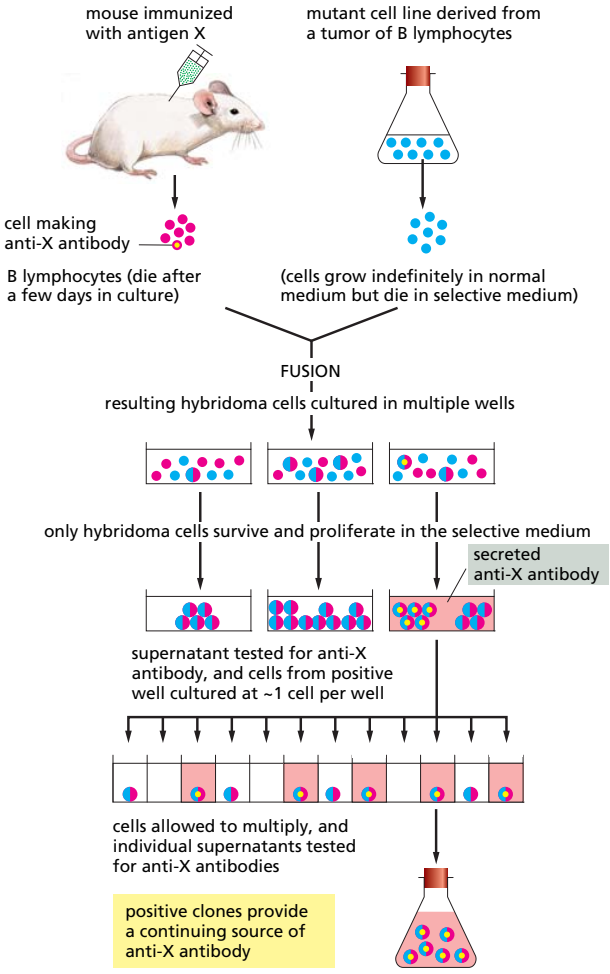



**Figure 8–8 Preparation of hybridomas that secrete monoclonal antibodies against a particular antigen.** Here, the antigen of interest is designated as "antigen X." The selective growth medium used after the cell fusion step contains an inhibitor (aminopterin) that blocks the normal biosynthetic pathways by which nucleotides are made. The cells must therefore use a bypass pathway to synthesize their nucleic acids. This pathway is defective in the mutant cell line derived from the B cell tumor, but it is intact in the normal cells obtained from the immunized mouse. Because neither cell type used for the initial fusion can survive and proliferate on its own, only the hybridoma cells do so.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 167 of 240

## Summary

*Tissues can be dissociated into their component cells, from which individual cell types can be purified and used for biochemical analysis or for the establishment of cell cultures. Many animal and plant cells survive and proliferate in a culture dish if they are provided with a suitable culture medium containing nutrients and appropriate signal molecules. Although many animal cells stop dividing after a finite number of cell divisions, cells that have been immortalized through spontaneous mutations or genetic manipulation can be maintained indefinitely as cell lines. Embryonic stem cells can proliferate indefinitely in a culture dish, while retaining the ability to differentiate into all the different cell types of the body. They therefore hold great medical promise. Hybridoma cells are widely employed to produce unlimited quantities of uniform monoclonal antibodies, which are used to detect and purify cell proteins, as well as to diagnose and treat diseases.*


# PURIFYING PROTEINS

The challenge of isolating a single type of protein from the thousands of other proteins present in a cell is a formidable one, but must be overcome in order to study protein function *in vitro*. As we shall see later in this chapter, *recombinant DNA technology* can enormously simplify this task by "tricking" cells into producing large quantities of a given protein, thereby making its purification much easier. Whether the source of the protein is an engineered cell or a natural tissue, a purification procedure usually starts with subcellular fractionation to reduce the complexity of the material, and is then followed by purification steps of increasing specificity.

## Cells Can Be Separated into Their Component Fractions

In order to purify a protein, it must first be extracted from inside the cell. Cells can be broken up in various ways: they can be subjected to osmotic shock or ultrasonic vibration, forced through a small orifice, or ground up in a blender. These procedures break many of the membranes of the cell (including the plasma membrane and endoplasmic reticulum) into fragments that immediately reseal to form small closed vesicles. If carefully carried out, however, the disruption procedures leave organelles such as nuclei, mitochondria, the Golgi apparatus, lysosomes, and peroxisomes largely intact. The suspension of cells is thereby reduced to a thick slurry (called a *homogenate* or *extract*) that contains a variety of membrane-enclosed organelles, each with a distinctive size, charge, and density. Provided that the homogenization medium has been carefully chosen (by trial and error for each organelle), the various components—including the vesicles derived from the endoplasmic reticulum, called microsomes—retain most of their original biochemical properties.

The different components of the homogenate must then be separated. Such cell fractionations became possible only after the commercial development in the early 1940s of an instrument known as the *preparative ultracentrifuge*, which rotates extracts of broken cells at high speeds (**Figure 8–9**). This treatment separates cell components by size and density: in general, the largest units experience the largest centrifugal force and move the most rapidly. At relatively low speed, large components such as nuclei sediment to form a pellet at the bottom of the centrifuge tube; at slightly higher speed, a pellet of mitochondria is deposited; and at even higher speeds and with longer periods of centrifugation, first the small closed vesicles and then the ribosomes can be collected (**Figure 8–10**). All of these fractions are impure, but many of the contaminants can be removed by resuspending the pellet and repeating the centrifugation procedure several times.

Centrifugation is the first step in most fractionations, but it separates only components that differ greatly in size. A finer degree of separation can be achieved by layering the homogenate in a thin band on top of a dilute salt solution that fills

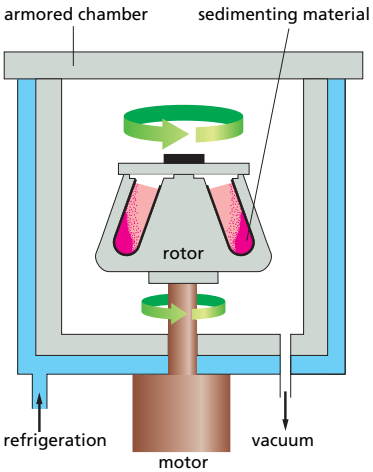



**Figure 8–9 The preparative ultracentrifuge.** The sample is contained in tubes that are inserted into a ring of cylindrical holes in a metal *rotor*. Rapid rotation of the rotor generates enormous centrifugal forces, which cause particles in the sample to sediment. The vacuum reduces friction, preventing heating of the rotor and allowing the refrigeration system to maintain the sample at 4°C.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 168 of 240

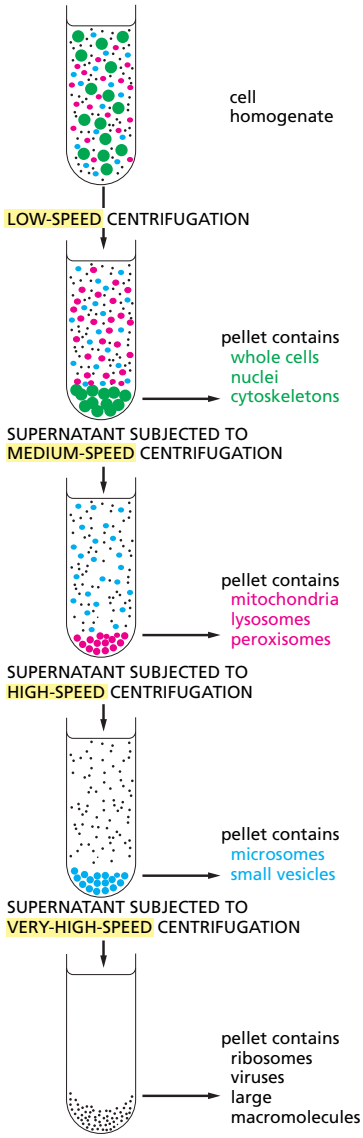



**Figure 8–10 Cell fractionation by centrifugation.** Repeated centrifugation at progressively higher speeds will fractionate homogenates of cells into their components. In general, the smaller the subcellular component, the greater is the centrifugal force required to sediment it. Typical values for the various centrifugation steps referred to in the figure are:
*low speed:* 1000 times gravity for 10 minutes
*medium speed:* 20,000 times gravity for 20 minutes
*high speed:* 80,000 times gravity for 1 hour
*very high speed:* 150,000 times gravity for 3 hours

a centrifuge tube. When centrifuged, the various components in the mixture move as a series of distinct bands through the salt solution, each at a different rate, in a process called *velocity sedimentation* (**Figure 8–11**A). For the procedure to work effectively, the bands must be protected from convective mixing, which would normally occur whenever a denser solution (for example, one containing organelles) finds itself on top of a lighter one (the salt solution). This is achieved by augmenting the solution in the tube with a shallow gradient of sucrose prepared by a special mixing device. The resulting density gradient—with the dense end at the bottom of the tube—keeps each region of the salt solution denser than any solution above it, and it thereby prevents convective mixing from distorting the separation.

When sedimented through such dilute sucrose gradients, different cell components separate into distinct bands that can be collected individually. The relative rate at which each component sediments depends primarily on its size and shape—normally being described in terms of its sedimentation coefficient, or S value. Present-day ultracentrifuges rotate at speeds of up to 80,000 rpm and produce forces as high as 500,000 times gravity. These enormous forces drive even small macromolecules, such as tRNA molecules and simple enzymes, to sediment at an appreciable rate and allow them to be separated from one another by size.

The ultracentrifuge is also used to separate cell components on the basis of their buoyant density, independently of their size and shape. In this case the sample is sedimented through a steep density gradient that contains a very high concentration of sucrose or cesium chloride. Each cell component begins to move down the gradient as in Figure 8–11A, but it eventually reaches a position where the density of the solution is equal to its own density. At this point the component floats and can move no farther. A series of distinct bands is thereby produced in the centrifuge tube, with the bands closest to the bottom of the tube containing the components of highest buoyant density (Figure 8–11B). This method, called *equilibrium sedimentation*, is so sensitive that it can separate macromolecules that have incorporated heavy isotopes, such as $^{13}C$ or $^{15}N$, from the same macromolecules that contain the lighter, common isotopes ($^{12}C$ or $^{14}N$). In fact, the cesium-chloride method was developed in 1957 to separate the labeled from the unlabeled DNA produced after exposure of a growing population of bacteria to nucleotide precursors containing $^{15}N$; this classic experiment provided direct evidence for the semiconservative replication of DNA (see Figure 5–5).

## Cell Extracts Provide Accessible Systems to Study Cell Functions

Studies of organelles and other large subcellular components isolated in the ultracentrifuge have contributed enormously to our understanding of the functions of different cell components. Experiments on mitochondria and chloroplasts purified by centrifugation, for example, demonstrated the central function of these organelles in converting energy into forms that the cell can use. Similarly, resealed vesicles formed from fragments of rough and smooth endoplasmic reticulum (microsomes) have been separated from each other and analyzed as functional models of these compartments of the intact cell.

Similarly, highly concentrated cell extracts, especially extracts of *Xenopus laevis* (African clawed frog) oocytes, have played a critical role in the study of

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 169 of 240

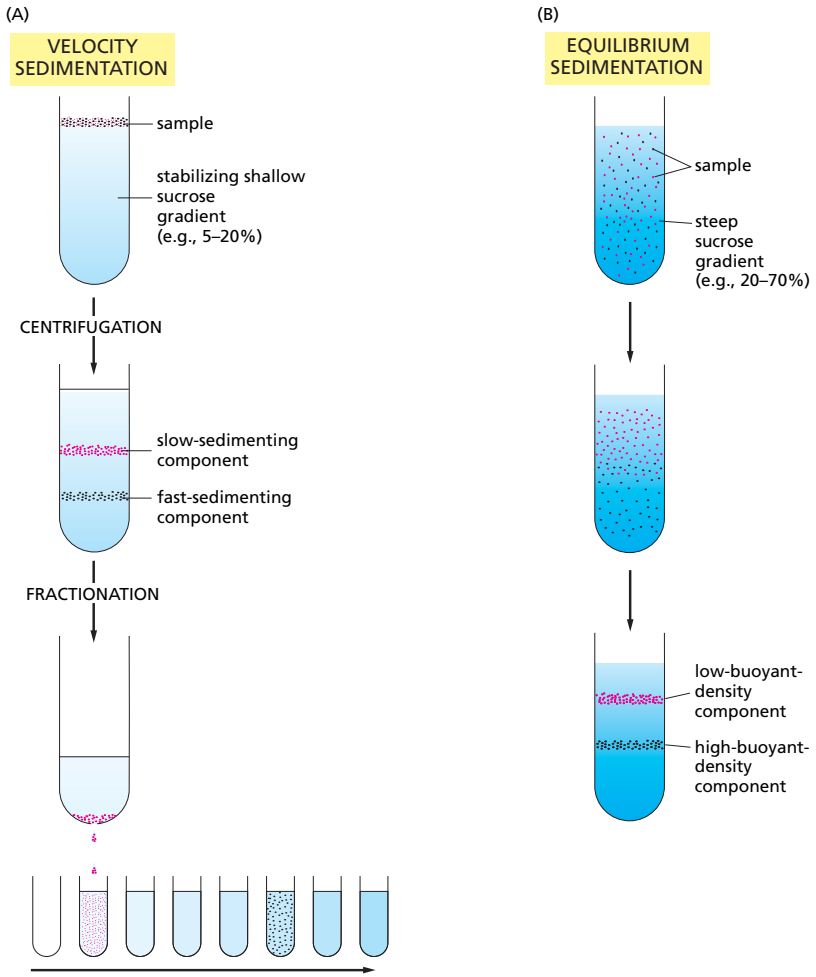



**Figure 8–11 Comparison of velocity sedimentation and equilibrium sedimentation.** (A) In velocity sedimentation, subcellular components sediment at different speeds according to their size and shape when layered over a dilute solution containing sucrose. To stabilize the sedimenting bands against convective mixing caused by small differences in temperature or solute concentration, the tube contains a continuous shallow gradient of sucrose, which increases in concentration toward the bottom of the tube (typically from 5% to 20% sucrose). After centrifugation, the different components can be collected individually, most simply by puncturing the plastic centrifuge tube and collecting drops from the bottom, as illustrated here. (B) In equilibrium sedimentation, subcellular components move up or down when centrifuged in a gradient until they reach a position where their density matches their surroundings. Although a sucrose gradient is shown here, denser gradients, which are especially useful for protein and nucleic acid separation, can be formed from cesium chloride. The final bands, at equilibrium, can be collected as in (A).

such complex and highly organized processes as the cell-division cycle, the separation of chromosomes on the mitotic spindle, and the vesicular-transport steps involved in the movement of proteins from the endoplasmic reticulum through the Golgi apparatus to the plasma membrane.

Cell extracts also provide, in principle, the starting material for the complete separation of all of the individual macromolecular components of the cell. We now consider how this separation is achieved, focusing on proteins.

## Proteins Can Be Separated by Chromatography

Proteins are most often fractionated by **column chromatography**, in which a mixture of proteins in solution is passed through a column containing a porous solid matrix. The different proteins are retarded to different extents by their interaction with the matrix, and they can be collected separately as they flow out of the bottom of the column (**Figure 8–12**). Depending on the choice of matrix, proteins can be separated according to their charge *(ion-exchange chromatography)*, their hydrophobicity *(hydrophobic chromatography)*, their size *(gel-filtration chromatography)*, or their ability to bind to particular small molecules or to other macromolecules *(affinity chromatography)*.

Many types of matrices are commercially available (**Figure 8–13**). Ion-exchange columns are packed with small beads that carry either a positive or a negative charge, so that proteins are fractionated according to the arrangement of charges on their surface. Hydrophobic columns are packed with beads from which hydrophobic side chains protrude, selectively retarding proteins with

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 170 of 240

exposed hydrophobic regions. Gel-filtration columns, which separate proteins according to their size, are packed with tiny porous beads: molecules that are small enough to enter the pores linger inside successive beads as they pass down the column, while larger molecules remain in the solution flowing between the beads and therefore move more rapidly, emerging from the column first. Besides providing a means of separating molecules, gel-filtration chromatography is a convenient way to determine their size.

Inhomogeneities in the matrices (such as cellulose), which cause an uneven flow of solvent through the column, limit the resolution of conventional column chromatography. Special chromatography resins (usually silica-based) composed of tiny spheres (3–10 μm in diameter) can be packed with a special apparatus to form a uniform column bed. Such **high-performance liquid chromatography** (**HPLC**) columns attain a high degree of resolution. In HPLC, the solutes equilibrate very rapidly with the interior of the tiny spheres, and so solutes with different affinities for the matrix are efficiently separated from one another even at very fast flow rates. HPLC is therefore the method of choice for separating many proteins and small molecules.

## Affinity Chromatography Exploits Specific Binding Sites on Proteins

If one starts with a complex mixture of proteins, the types of column chromatography just discussed do not produce very highly purified fractions: a single passage through the column generally increases the proportion of a given protein in the mixture no more than twentyfold. Because most individual proteins represent less than 1/1000 of the total cell protein, it is usually necessary to use several different types of columns in succession to attain sufficient purity (**Figure 8–14**). A more efficient procedure, known as **affinity chromatography**, takes advantage of the biologically important binding interactions that occur on protein surfaces. If a substrate molecule is covalently coupled to an inert matrix such as a polysaccharide bead, the enzyme that operates on that substrate will often be specifically retained by the matrix and can then be eluted (washed out) in nearly pure form. Likewise, short DNA oligonucleotides of a specifically

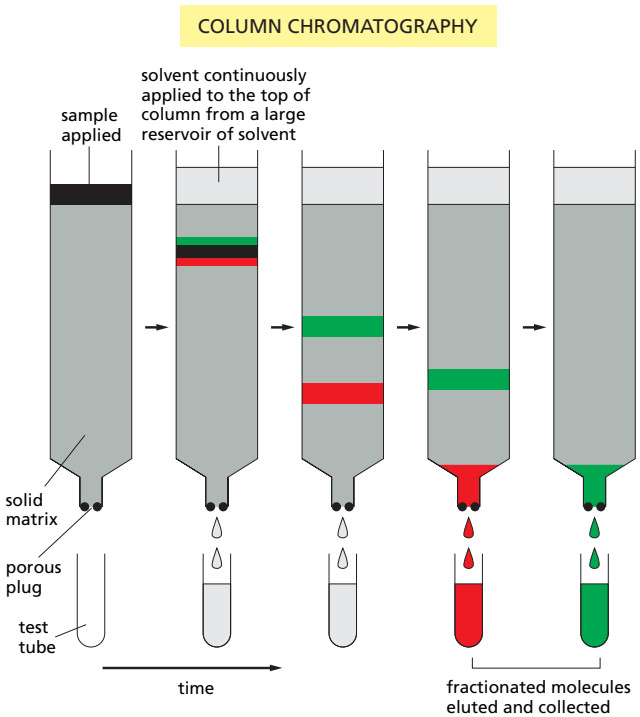



**Figure 8–12 The separation of molecules by column chromatography.** The sample, a solution containing a mixture of different molecules, is applied to the top of a cylindrical glass or plastic column filled with a permeable solid matrix, such as cellulose. A large amount of solvent is then pumped slowly through the column and collected in separate tubes as it emerges from the bottom. Because various components of the sample travel at different rates through the column, they are fractionated into different tubes.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 171 of 240

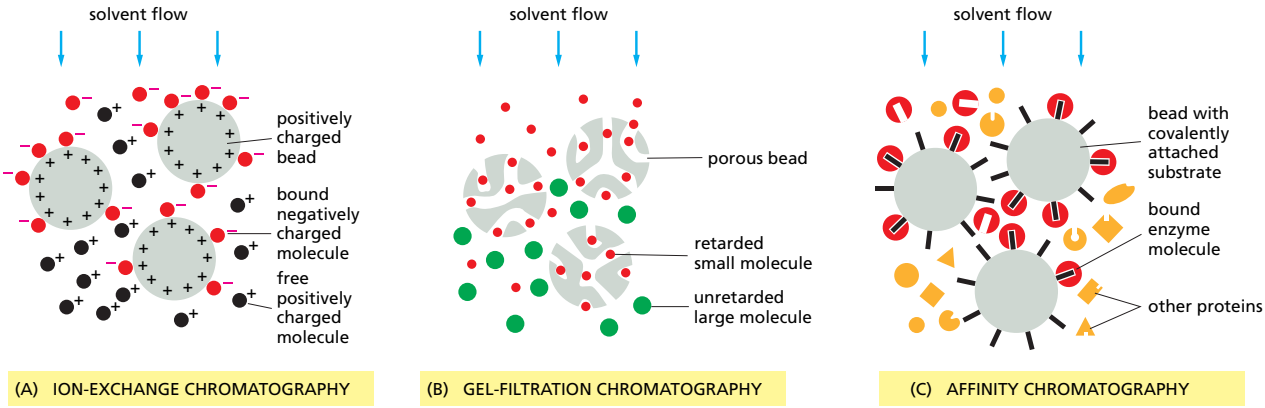



**Figure 8–13 Three types of matrices used for chromatography.** (A) In ion-exchange chromatography, the insoluble matrix carries ionic charges that retard the movement of molecules of opposite charge. Matrices used for separating proteins include diethylaminoethylcellulose (DEAE-cellulose), which is positively charged, and carboxymethylcellulose (CM-cellulose) and phosphocellulose, which are negatively charged. Analogous matrices based on agarose or other polymers are also frequently used. The strength of the association between the dissolved molecules and the ion-exchange matrix depends on both the ionic strength and the pH of the solution that is passing down the column, which may therefore be varied systematically (as in Figure 8–14) to achieve an effective separation. (B) In gel-filtration chromatography, the matrix is inert but porous. Molecules that are small enough to penetrate into the matrix are thereby delayed and travel more slowly through the column than larger molecules that cannot penetrate. Beads of cross-linked polysaccharide (dextran, agarose, or acrylamide) are available commercially in a wide range of pore sizes, making them suitable for the fractionation of molecules of various molecular weights, from less than 500 daltons to more than $5 \times 10^6$ daltons. (C) Affinity chromatography uses an insoluble matrix that is covalently linked to a specific ligand, such as an antibody molecule or an enzyme substrate, that will bind a specific protein. Enzyme molecules that bind to immobilized substrates on such columns can be eluted with a concentrated solution of the free form of the substrate molecule, while molecules that bind to immobilized antibodies can be eluted by dissociating the antibody–antigen complex with concentrated salt solutions or solutions of high or low pH. High degrees of purification can be achieved in a single pass through an affinity column.

designed sequence can be immobilized in this way and used to purify DNA-binding proteins that normally recognize this sequence of nucleotides in chromosomes (see Figure 7–28). Alternatively, specific antibodies can be coupled to a matrix to purify protein molecules recognized by the antibodies. Because of the great specificity of all such affinity columns, 1000- to 10,000-fold purifications can sometimes be achieved in a single pass.

## Genetically-Engineered Tags Provide an Easy Way to Purify Proteins

Using the recombinant DNA methods discussed in subsequent sections, any gene can be modified to produce its protein with a special recognition tag attached to it, so as to make subsequent purification of the protein by affinity chromatography simple and rapid. Often the recognition tag is itself an antigenic determinant, or *epitope*, which can be recognized by a highly specific antibody. The antibody, can then be used both to localize the protein in cells and to purify it (**Figure 8–15**). Other types of tags are specifically designed for protein purification. For example, the amino acid histidine binds to certain metal ions, including nickel and copper. If genetic engineering techniques are used to attach a short string of histidines to one end of a protein, the slightly modified protein can be retained selectively on an affinity column containing immobilized nickel ions. Metal affinity chromatography can thereby be used to purify the modified protein from a complex molecular mixture.

In other cases, an entire protein is used as the recognition tag. When cells are engineered to synthesize the small enzyme glutathione S-transferase (GST) attached to a protein of interest, the resulting **fusion protein** can be purified from the other contents of the cell with an affinity column containing glutathione, a

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 172 of 240

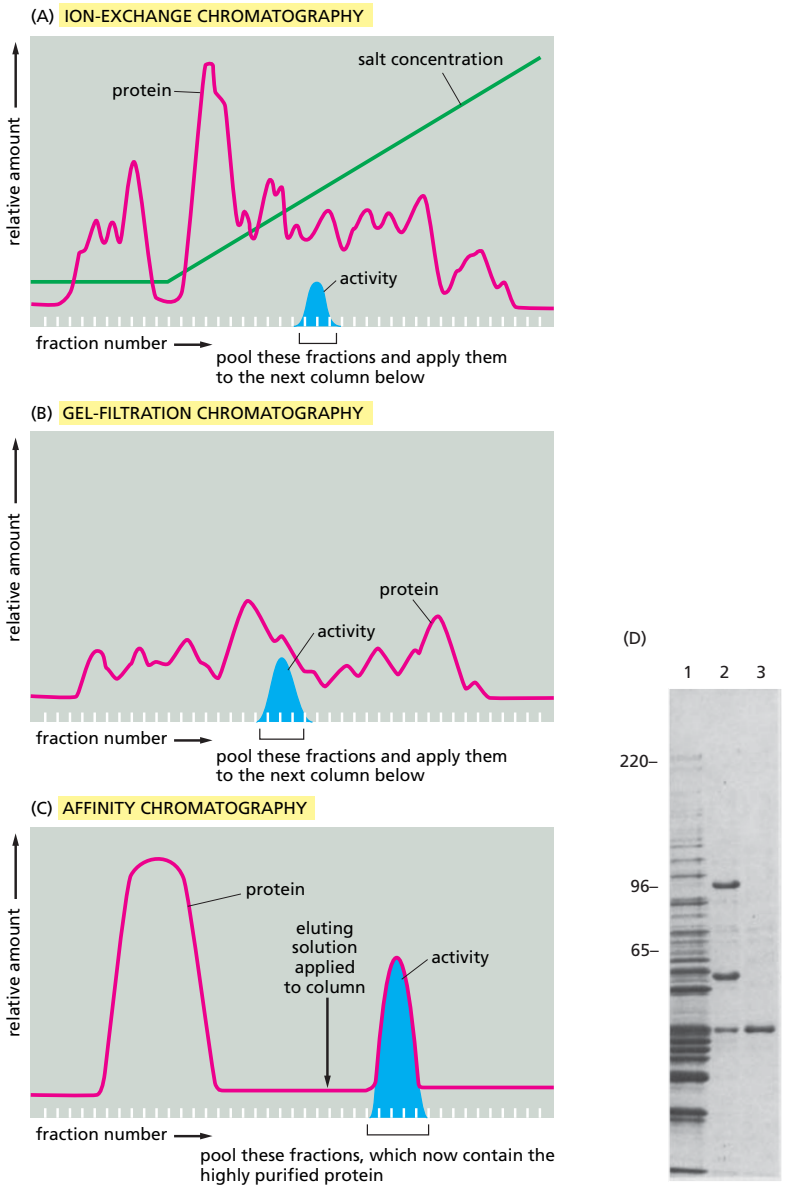



**Figure 8–14 Protein purification by chromatography.** Typical results obtained when three different chromatographic steps are used in succession to purify a protein. In this example, a homogenate of cells was first fractionated by allowing it to percolate through an ion-exchange resin packed into a column (A). The column was washed to remove all unbound contaminants, and the bound proteins were then eluted by passing a solution containing a gradually increasing concentration of salt onto the top of the column. Proteins with the lowest affinity for the ion-exchange resin passed directly through the column and were collected in the earliest fractions eluted from the bottom of the column. The remaining proteins were eluted in sequence according to their affinity for the resin—those proteins binding most tightly to the resin requiring the highest concentration of salt to remove them. The protein of interest was eluted in several fractions and was detected by its enzymatic activity. The fractions with activity were pooled and then applied to a second, gel-filtration column (B). The elution position of the still-impure protein was again determined by its enzymatic activity, and the active fractions were pooled and purified to homogeneity on an affinity column (C) that contained an immobilized substrate of the enzyme. (D) Affinity purification of cyclin-binding proteins from *S. cerevisiae*, as analyzed by SDS polyacrylamide-gel electrophoresis, which is described below in Figure 8–18. Lane 1 is a total cell extract; lane 2 shows the proteins eluted from an affinity column containing cyclin B2; lane 3 shows one major protein eluted from a cyclin B3 affinity column. Proteins in lanes 2 and 3 were eluted from the affinity columns with salt, and the gel was stained with Coomassie blue. The scale at the left shows the molecular weights of marker proteins, in kilodaltons. (D, from D. Kellogg et al., *J. Cell Biol.* 130:675–685, 1995. With permisison from The Rockefeller University Press.)

substrate molecule that binds specifically and tightly to GST. If the purification is carried out under conditions that do not disrupt protein–protein interactions, the fusion protein can be isolated in association with the proteins it interacts with inside the cell (**Figure 8–16**).

As a further refinement of purification methods using recognition tags, an amino acid sequence that forms a cleavage site for a highly specific proteolytic enzyme can be engineered between the protein of choice and the recognition tag. Because the amino acid sequences at the cleavage site are very rarely found by chance in proteins, the tag can later be cleaved off without destroying the purified protein.

This type of specific cleavage is used in an especially powerful purification methodology known as *tandem affinity purification tagging (tap-tagging)*. Here, one end of a protein is engineered to contain two recognition tags that are separated by a protease cleavage site. The tag on the very end of the construct is chosen to bind irreversibly to an affinity column, allowing the column to be washed extensively to remove all contaminating proteins. Protease cleavage then releases the protein, which is then further purified using the second tag.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 173 of 240

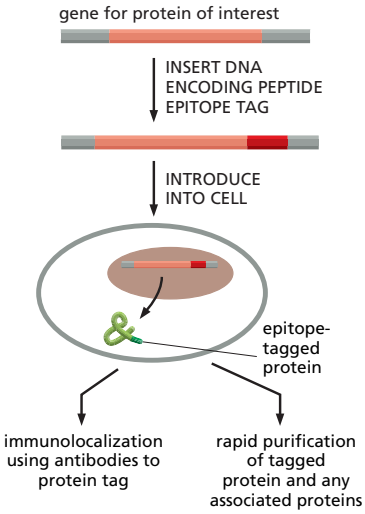



**Figure 8–15 Epitope tagging for the localization or purification of proteins.** Using standard genetic engineering techniques, a short peptide tag can be added to a protein of interest. If the tag is itself an antigenic determinant, or *epitope*, it can be targeted by an appropriate commercially available antibody. The antibody, suitably labeled, can be used to determine the location of the protein in cells or to purify it by immunoprecipitation or affinity chromatography. In immunoprecipitation, antibodies directed against the epitope tag are added to a solution containing the tagged protein; the antibodies specifically cross-link the tagged protein molecules and precipitate them out of solution as antibody–protein complexes.

Because this two-step strategy provides an especially high degree of protein purification with relatively little effort, it is used extensively in cell biology. Thus, for example, a set of approximately 6000 yeast strains, each with a different gene fused to DNA that encodes a tap-tag, has been constructed to allow any yeast protein to be rapidly purified.

## Purified Cell-free Systems Are Required for the Precise Dissection of Molecular Functions

It is important to study biological processes free from all of the complex side reactions that occur in a living cell by using **purified cell-free systems**. To make this possible, cell homogenates are fractionated with the aim of purifying each of the individual macromolecules that are needed to catalyze a biological process of interest. For example, the experiments to decipher the mechanisms of protein synthesis began with a cell homogenate that could translate RNA molecules to produce proteins. Fractionation of this homogenate, step by step, produced in turn the ribosomes, tRNAs, and various enzymes that together constitute the protein-synthetic machinery. Once individual pure components were available, each could be added or withheld separately to define its exact role in the overall process.

A major goal for cell biologists is the reconstitution of every biological process in a purified cell-free system. Only in this way can one define all of the components needed for the process and control their concentrations, as required to work out their precise mechanism of action. Although much remains to be done, a great deal of what we know today about the molecular biology of the cell has been discovered by studies in such cell-free systems. They have been used, for example, to decipher the molecular details of DNA replication and DNA transcription, RNA splicing, protein translation, muscle contraction, and particle transport along microtubules, and many other processes that occur in cells.

## Summary

*Populations of cells can be analyzed biochemically by disrupting them and fractionating their contents, allowing functional cell-free systems to be developed. Highly purified cell-free systems are needed for determining the molecular details of complex cell processes, requiring extensive purification of all the proteins and other components involved. The proteins in soluble cell extracts can be purified by column*

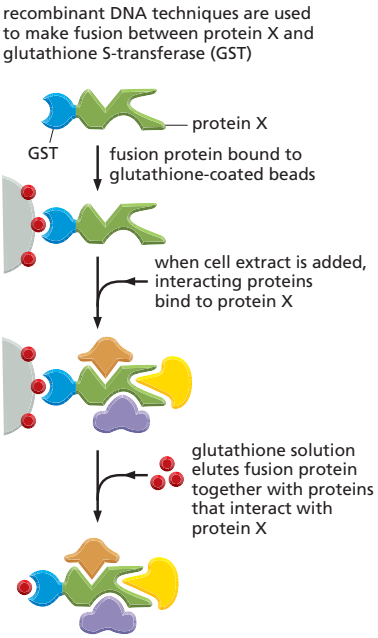



**Figure 8–16 Purification of protein complexes by using a GST-tagged fusion protein.** GST fusion proteins, produced in cells with recombinant DNA techniques, can be captured on an affinity column containing beads coated with glutathione. Proteins not bound to the beads are washed away. The fusion protein, along with other proteins in the cell that are bound tightly to it, can then be eluted with glutathione. The identities of these additional proteins can be determined by mass spectrometry (see Figure 8–21). Affinity columns can also be made to contain antibodies against GST or another convenient small protein or epitope tag (see Figure 8–15).

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 174 of 240

*chromatography; depending on the type of column matrix, biologically active proteins can be separated on the basis of their molecular weight, hydrophobicity, charge characteristics, or affinity for other molecules. In a typical purification, the sample is passed through several different columns in turn—the enriched fractions obtained from one column are applied to the next. Recombinant DNA techniques, to be described later, allow special recognition tags to be attached to proteins, thereby greatly simplifying their purification.*

# ANALYZING PROTEINS

Proteins perform most processes in cells: they catalyze metabolic reactions, use nucleotide hydrolysis to do mechanical work, and serve as the major structural elements of the cell. The great variety of protein structures and functions has stimulated the development of a multitude of techniques to study them.

## Proteins Can Be Separated by SDS Polyacrylamide-Gel Electrophoresis

Proteins usually possess a net positive or negative charge, depending on the mixture of charged amino acids they contain. An electric field applied to a solution containing a protein molecule causes the protein to migrate at a rate that depends on its net charge and on its size and shape. The most popular application of this property is **SDS polyacrylamide-gel electrophoresis** (**SDS-PAGE**). It uses a highly cross-linked gel of polyacrylamide as the inert matrix through which the proteins migrate. The gel is prepared by polymerization of monomers; the pore size of the gel can be adjusted so that it is small enough to retard the migration of the protein molecules of interest. The proteins themselves are not in a simple aqueous solution but in one that includes a powerful negatively charged detergent, sodium dodecyl sulfate, or SDS (**Figure 8–17**). Because this detergent binds to hydrophobic regions of the protein molecules, causing them to unfold into extended polypeptide chains, the individual protein molecules are released from their associations with other proteins or lipid molecules and rendered freely soluble in the detergent solution. In addition, a reducing agent such as β-mercaptoethanol (see Figure 8–17) is usually added to break any S–S linkages in the proteins, so that all of the constituent polypeptides in multisubunit proteins can be analyzed separately.

What happens when a mixture of SDS-solubilized proteins is run through a slab of polyacrylamide gel? Each protein molecule binds large numbers of the negatively charged detergent molecules, which mask the protein's intrinsic charge and cause it to migrate toward the positive electrode when a voltage is applied. Proteins of the same size tend to move through the gel with similar speeds because (1) their native structure is completely unfolded by the SDS, so that their shapes are the same, and (2) they bind the same amount of SDS and therefore have the same amount of negative charge. Larger proteins, with more charge, are subjected to larger electrical forces and also to a larger drag. In free solution, the two effects would cancel out, but, in the mesh of the polyacrylamide gel, which acts as a molecular sieve, large proteins are retarded much more than small ones. As a result, a complex mixture of proteins is fractionated into a series of discrete protein bands arranged in order of molecular weight (**Figure 8–18**). The major proteins are readily detected by staining the proteins in the gel with a dye such as Coomassie blue. Even minor proteins are seen in gels treated with a silver or gold stain, so that as little as 10 ng of protein can be detected in a band.

SDS-PAGE is widely used because it can separate all types of proteins, including those that are normally insoluble in water—such as the many proteins in membranes. And because the method separates polypeptides by size, it provides information about the molecular weight and the subunit composition of proteins. **Figure 8–19** presents a photograph of a gel that has been used to analyze each of the successive stages in the purification of a protein.

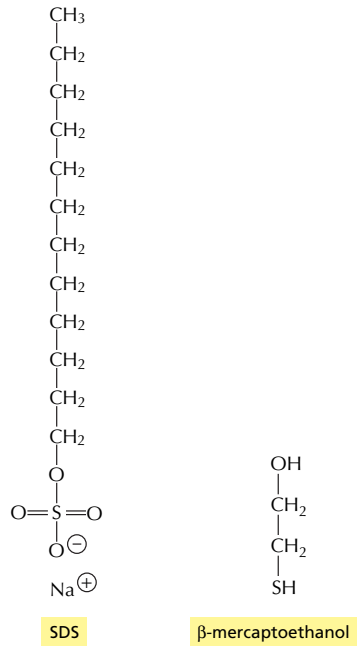



**Figure 8–17 The detergent sodium dodecyl sulfate (SDS) and the reducing agent β-mercaptoethanol.** These two chemicals are used to solubilize proteins for SDS polyacrylamide-gel electrophoresis. The SDS is shown here in its ionized form.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 175 of 240

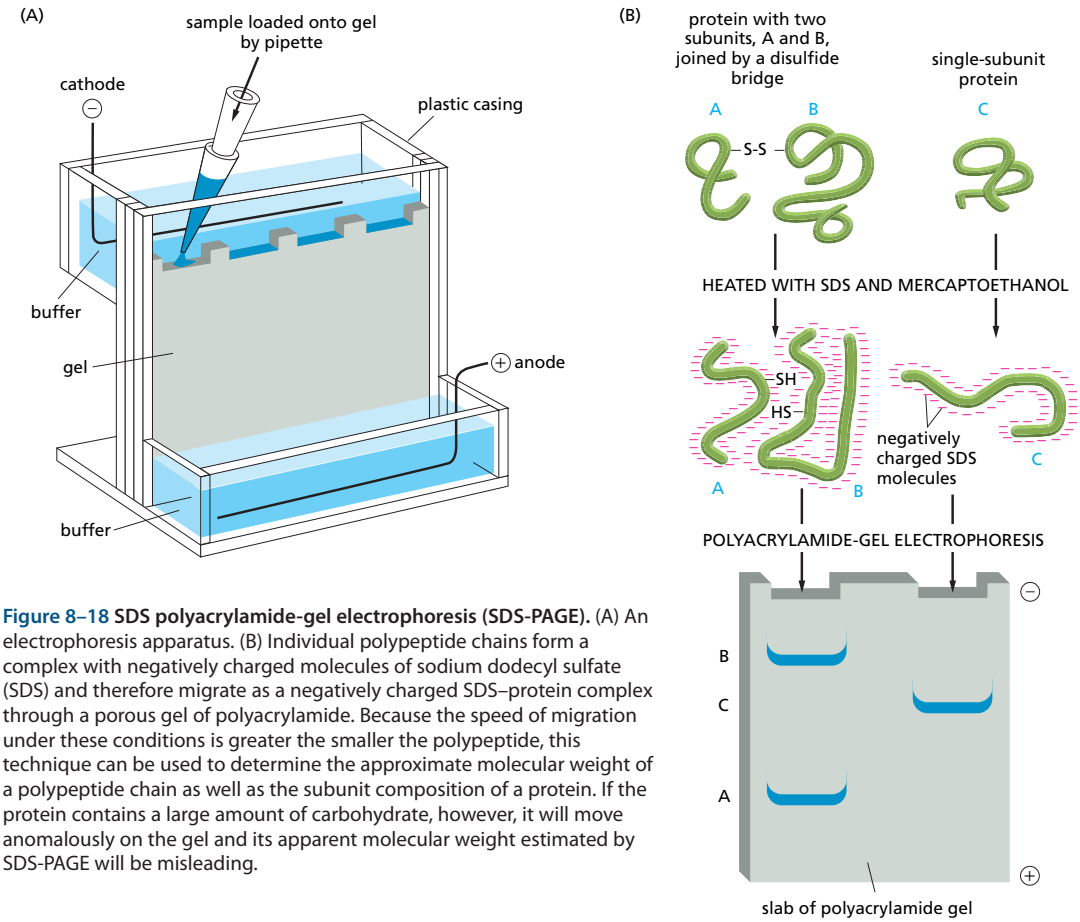



**Figure 8–18 SDS polyacrylamide-gel electrophoresis (SDS-PAGE).** (A) An electrophoresis apparatus. (B) Individual polypeptide chains form a complex with negatively charged molecules of sodium dodecyl sulfate (SDS) and therefore migrate as a negatively charged SDS–protein complex through a porous gel of polyacrylamide. Because the speed of migration under these conditions is greater the smaller the polypeptide, this technique can be used to determine the approximate molecular weight of a polypeptide chain as well as the subunit composition of a protein. If the protein contains a large amount of carbohydrate, however, it will move anomalously on the gel and its apparent molecular weight estimated by SDS-PAGE will be misleading.

## Specific Proteins Can Be Detected by Blotting with Antibodies

A specific protein can be identified after its fractionation on a polyacrylamide gel by exposing all the proteins present on the gel to a specific antibody that has been coupled to a radioactive isotope, to an easily detectable enzyme, or to a fluorescent dye. For convenience, this procedure is normally carried out after transferring (by "blotting") all of the separated proteins present in the gel onto a sheet of nitrocellulose paper or nylon membrane. Placing the membrane over the gel and driving the proteins out of the gel with a strong electric field transfers the protein onto the membrane. The membrane is then soaked in a solution of labeled antibody to reveal the protein of interest. This method of detecting proteins is called **Western blotting**, or **immunoblotting** (**Figure 8–20**).

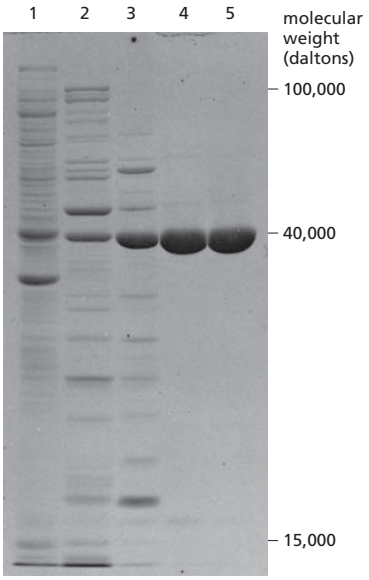
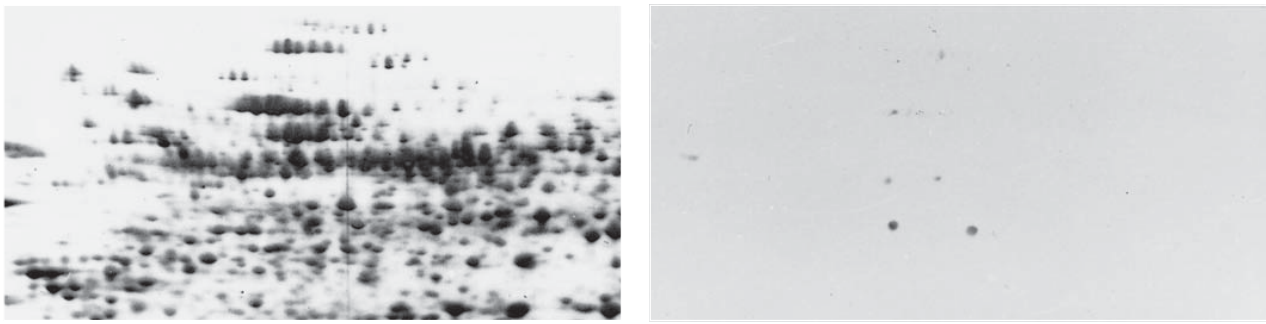



**Figure 8–19 Analysis of protein samples by SDS polyacrylamide-gel electrophoresis.** The photograph shows a Coomassie-stained gel that has been used to detect the proteins present at successive stages in the purification of an enzyme. The leftmost lane (lane 1) contains the complex mixture of proteins in the starting cell extract, and each succeeding lane analyzes the proteins obtained after a chromatographic fractionation of the protein sample analyzed in the previous lane (see Figure 8–14). The same total amount of protein (10 μg) was loaded onto the gel at the top of each lane. Individual proteins normally appear as sharp, dye-stained bands; a band broadens, however, when it contains too much protein. (From T. Formosa and B.M. Alberts, *J. Biol. Chem.* 261:6107–6118, 1986.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 176 of 240



(A) (B)

**Figure 8–20 Western blotting.** All the proteins from dividing tobacco cells in culture are first separated by two-dimensional polyacrylamide-gel electrophoresis (described in Figure 8–23). In (A), the positions of the proteins are revealed by a sensitive protein stain. In (B), the separated proteins on an identical gel were then transferred to a sheet of nitrocellulose and exposed to an antibody that recognizes only those proteins that are phosphorylated on threonine residues during mitosis. The positions of the dozen or so proteins that are recognized by this antibody are revealed by an enzyme-linked second antibody. This technique is also known as immunoblotting (or Western blotting). (From J.A. Traas et al., *Plant J.* 2:723–732, 1992. With permission from Blackwell Publishing.)

## Mass Spectrometry Provides a Highly Sensitive Method for Identifying Unknown Proteins

A frequent problem in cell biology and biochemistry is the identification of a protein or collection of proteins that has been obtained by one of the purification procedures discussed in the preceding pages (see, for example, Figure 8–16). Because the genome sequences of most common experimental organisms are now known, catalogues of all the proteins produced in those organisms are available. The task of identifying an unknown protein (or collection of unknown proteins) thus reduces to matching some of the amino acid sequences present in the unknown sample with known catalogued genes. This task is now performed almost exclusively by using mass spectrometry in conjunction with computer searches of databases.

Charged particles have very precise dynamics when subjected to electrical and magnetic fields in a vacuum. Mass spectrometry exploits this principle to separate ions according to their mass-to-charge ratio. It is an enormously sensitive technique. It requires very little material and is capable of determining the precise mass of intact proteins and of peptides derived from them by enzymatic or chemical cleavage. Masses can be obtained with great accuracy, often with an error of less than one part in a million. The most commonly used form of the technique is called *matrix-assisted laser desorption ionization–time-of-flight spectrometry (MALDI-TOF)*. In this approach, the proteins in the sample are first broken into short peptides. These peptides are mixed with an organic acid and then dried onto a metal or ceramic slide. A laser then blasts the sample, ejecting the peptides from the slide in the form of an ionized gas, in which each molecule carries one or more positive charges. The ionized peptides are accelerated in an electric field and fly toward a detector. Their mass and charge determines the time it takes them to reach the detector: large peptides move more slowly, and more highly charged molecules move more quickly. By analyzing those ionized peptides that bear a single charge, the precise masses of peptides present in the original sample can be determined. MALDI-TOF can also be used to accurately measure the mass of intact proteins as large as 200,000 daltons. This information is then used to search genomic databases, in which the masses of all proteins and of all their predicted peptide fragments have been tabulated from the genomic sequences of the organism (**Figure 8–21**A). An unambiguous match to a particular open reading frame can often be made by knowing the mass of only a few peptides derived from a given protein.

MALDI-TOF provides accurate molecular weight measurements for proteins and peptides. Moreover, by employing two mass spectrometers in tandem (an arrangement known as MS/MS), it is possible to directly determine the

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 177 of 240

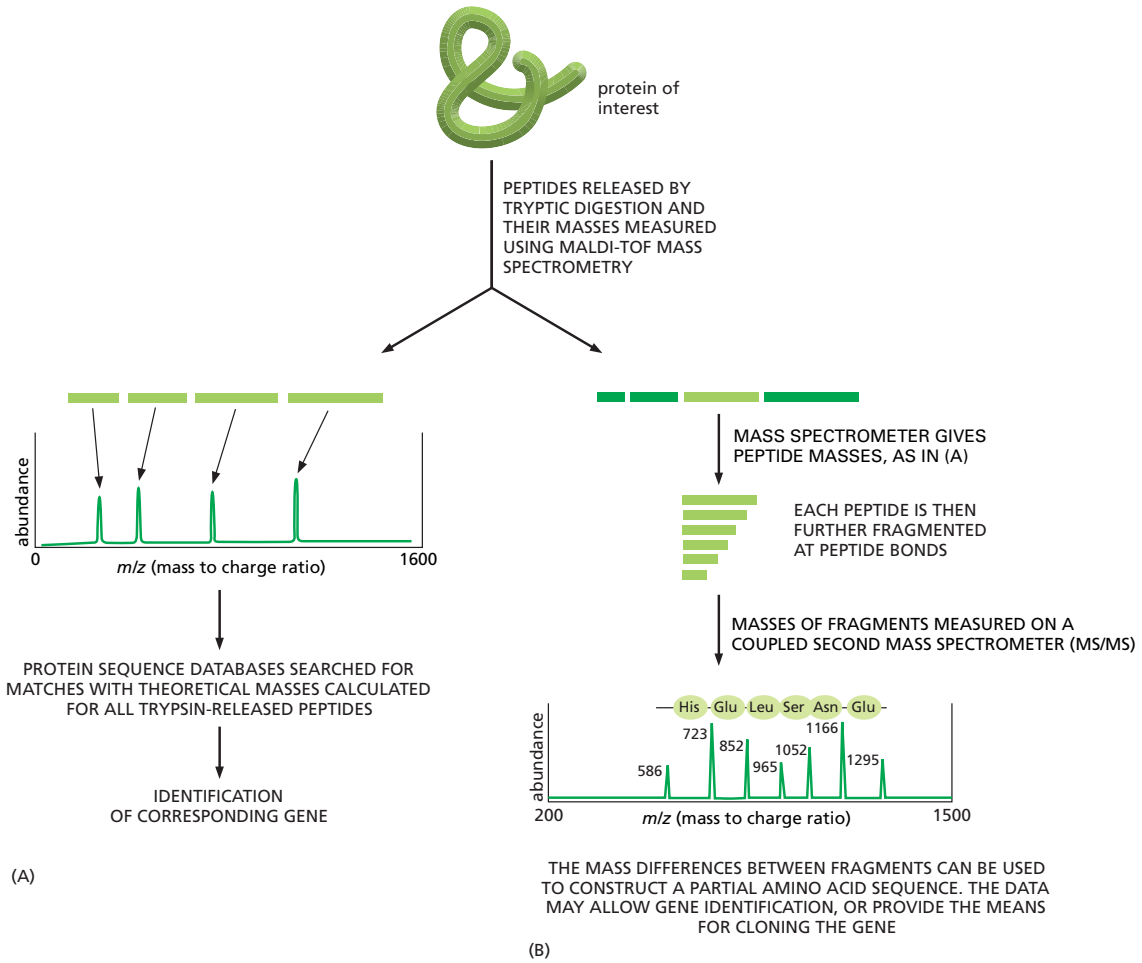



**Figure 8–21 Use of mass spectrometry to identify proteins and to sequence peptides.** An isolated protein is digested with trypsin and the peptide fragments are then loaded into the mass spectrometer. Two different approaches can then be used to identify the protein. (A) In the first method, peptide masses are measured precisely using MALDI-TOF mass spectrometry. Sequence databases are then searched to find the gene that encodes a protein whose calculated tryptic digest profile matches these values. (B) Mass spectrometry can also be used to determine directly the amino acid sequence of peptide fragments. In this example, tryptic peptides are first separated based on mass within a mass spectrometer. Each peptide is then further fragmented, primarily by cleaving its peptide bonds. This treatment generates a nested set of peptides, each differing in size by one amino acid. These fragments are fed into a second coupled mass spectrometer, and their masses are determined. The difference in masses between two closely related peptides can be used to deduce the "missing" amino acid. By repeated applications of this procedure, a partial amino acid sequence of the original protein can be determined. For simplicity, the analysis shown begins with a single species of purified protein. In reality, mass spectrometry is usually carried out on mixtures of proteins, such as those obtained for affinity chromatography experiments (see Figure 8–16), and can identify all the proteins present in the mixtures. As explained in the text, mass spectrometry can also detect post-translational modifications of proteins.

amino acid sequences of individual peptides in a complex mixture. As described above, the protein sample is first broken down into smaller peptides, which are separated from each other by mass spectrometry. Each peptide is then further fragmented through collisions with high-energy gas atoms. This method of fragmentation preferentially cleaves the peptide bonds, generating a ladder of fragments, each differing by a single amino acid. The second mass spectrometer then separates these fragments and displays their masses. The amino acid sequence of a peptide can then be deduced from these differences in mass (Figure 8–21B).

MS/MS is particularly useful for detecting and precisely mapping post-translational modifications of proteins, such as phosphorylations or acetylations. Because these modifications impart a characteristic mass increase to an amino acid, they are easily detected by mass spectrometry. As described in

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 178 of 240

Chapter 3, proteomics, a general term that encompasses many different experimental techniques, is the characterization of all proteins in the cell, including all protein–protein interactions and all post-translational modifications. In combination with the rapid purification techniques discussed in the last section, mass spectrometry has emerged as the most powerful method for mapping both the post-translational modifications of a given protein and the proteins that remain associated with it during purification.

## Two-Dimensional Separation Methods are Especially Powerful

Because different proteins can have similar sizes, shapes, masses, and overall charges, most separation techniques such as SDS polyacrylamide-gel electrophoresis or ion-exchange chromatography cannot typically display all the proteins in a cell or even in an organelle. In contrast, **two-dimensional gel electrophoresis**, which combines two different separation procedures, can resolve up to 2000 proteins—the total number of different proteins in a simple bacterium—in the form of a two-dimensional protein map.

In the first step, the proteins are separated by their intrinsic charges. The sample is dissolved in a small volume of a solution containing a nonionic (uncharged) detergent, together with β-mercaptoethanol and the denaturing reagent urea. This solution solubilizes, denatures, and dissociates all the polypeptide chains but leaves their intrinsic charge unchanged. The polypeptide chains are then separated in a pH gradient by a procedure called *isoelectric focusing,* which takes advantage of the variation in the net charge on a protein molecule with the pH of its surrounding solution. Every protein has a characteristic isoelectric point, the pH at which the protein has no net charge and therefore does not migrate in an electric field. In isoelectric focusing, proteins are separated electrophoretically in a narrow tube of polyacrylamide gel in which a gradient of pH is established by a mixture of special buffers. Each protein moves to a position in the gradient that corresponds to its isoelectric point and remains there (**Figure 8–22**). This is the first dimension of two-dimensional polyacrylamide-gel electrophoresis.

In the second step, the narrow gel containing the separated proteins is again subjected to electrophoresis but in a direction that is at a right angle to the direction used in the first step. This time SDS is added, and the proteins separate according to their size, as in one-dimensional SDS-PAGE: the original narrow gel is soaked in SDS and then placed on one edge of an SDS polyacrylamide-gel slab, through which each polypeptide chain migrates to form a discrete spot. This is the second dimension of two-dimensional polyacrylamide-gel electrophoresis. The only proteins left unresolved are those that have both identical sizes and identical isoelectric points, a relatively rare situation. Even trace amounts of each polypeptide chain can be detected on the gel by various staining procedures—or by autoradiography if the protein sample was initially labeled with a radioisotope (**Figure 8–23**). The technique has such great resolving power that it can distinguish between two proteins that differ in only a single charged amino acid.

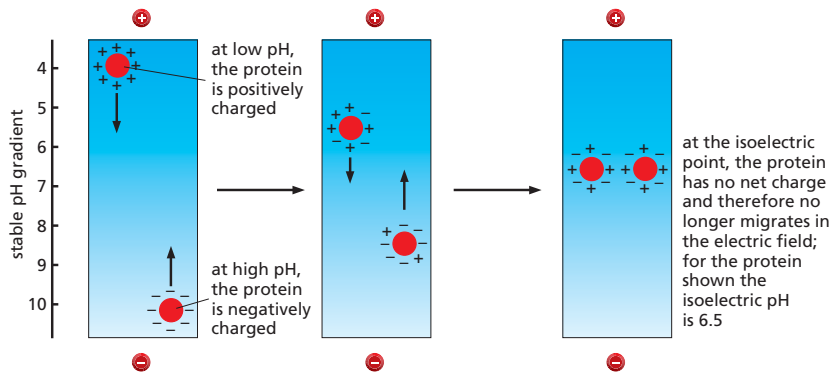



**Figure 8–22 Separation of protein molecules by isoelectric focusing.** At low pH (high $H^+$ concentration), the carboxylic acid groups of proteins tend to be uncharged (–COOH) and their nitrogen-containing basic groups fully charged (for example, $–NH_3^+$), giving most proteins a net positive charge. At high pH, the carboxylic acid groups are negatively charged ($–COO^-$) and the basic groups tend to be uncharged (for example, $–NH_2$), giving most proteins a net negative charge. At its *isoelectric pH,* a protein has no net charge since the positive and negative charges balance. Thus, when a tube containing a fixed pH gradient is subjected to a strong electric field in the appropriate direction, each protein species present migrates until it forms a sharp band at its isoelectric pH, as shown.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 179 of 240

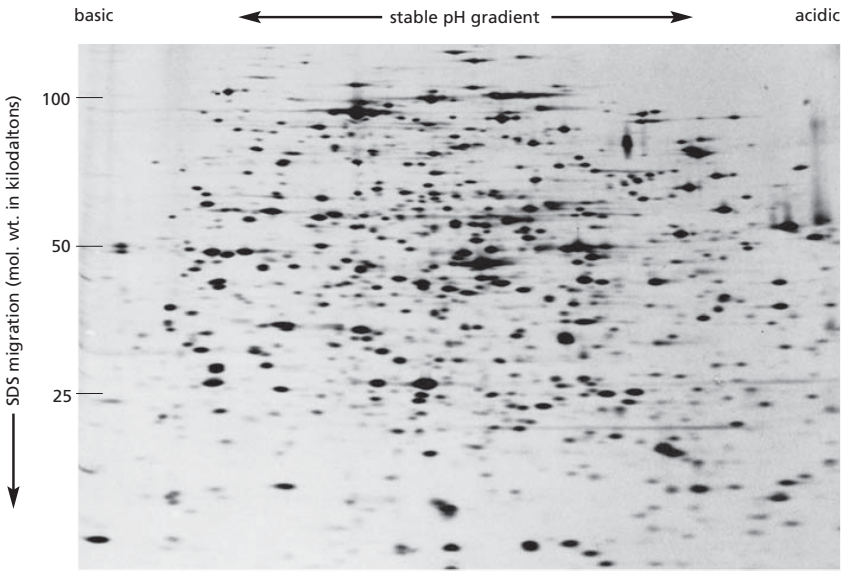



**Figure 8–23 Two-dimensional polyacrylamide-gel electrophoresis.** All the proteins in an *E. coli* bacterial cell are separated in this gel, in which each spot corresponds to a different polypeptide chain. The proteins were first separated on the basis of their isoelectric points by isoelectric focusing from left to right. They were then further fractionated according to their molecular weights by electrophoresis from top to bottom in the presence of SDS. Note that different proteins are present in very different amounts. The bacteria were fed with a mixture of radioisotope-labeled amino acids so that all of their proteins were radioactive and could be detected by autoradiography (see pp. 602–603). (Courtesy of Patrick O'Farrell.)

A different, even more powerful, "two-dimensional" technique is now available when the aim is to determine all of the proteins present in an organelle or another complex mixture of proteins. Because the technique relies on mass spectroscopy, it requires that the proteins be from an organism with a completely sequenced genome. First, the mixture of proteins present is digested with trypsin to produce short peptides. Next, these peptides are separated by a series of automated liquid chromatography steps. As the second dimension, each separated peptide is fed directly into a tandem mass spectrometer (MS/MS) that allows its amino acid sequence, as well as any post-translational modifications, to be determined. This arrangement, in which a tandem mass spectrometer (MS/MS) is attached to the output of an automated liquid chromatography (LC) system, is referred to as LC-MS/MS. It is now becoming routine to subject an entire organelle preparation to LC-MS/MS analysis and to identify hundreds of proteins and their modifications. Of course, no organelle isolation procedure is perfect, and some of the proteins identified will be contaminating proteins. These can often be excluded by analyzing neighboring fractions from the organelle purification and "subtracting" them out from the peak organelle fractions.

## Hydrodynamic Measurements Reveal the Size and Shape of a Protein Complex

Most proteins in a cell act as part of larger complexes, and knowledge of the size and shape of these complexes often leads to insights regarding their function. This information can be obtained in several important ways. Sometimes, a complex can be directly visualized using electron microscopy, as described in Chapter 9. A complementary approach relies on the hydrodynamic properties of a complex, that is, its behavior as it moves through a liquid medium. Usually, two separate measurements are made. One measure is the velocity of a complex as it moves under the influence of a centrifugal field produced by an ultracentrifuge (see Figure 8–11A). The sedimentation constant (or S-value) obtained depends on both the size and the shape of the complex and does not, by itself, convey especially useful information. However, once a second hydrodynamic measurement is performed—by charting the migration of a complex through a gel-filtration chromatography column (see Figure 8–13B)—both the approximate shape of a complex and its molecular weight can be calculated.

Molecular weight can also be determined more directly by using an analytical ultracentrifuge, a complex device that allows protein absorbance measurements

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 180 of 240

to be made on a sample while it is subjected to centrifugal forces. In this approach, the sample is centrifuged until it reaches equilibrium, where the centrifugal force on a protein complex exactly balances its tendency to diffuse away. Because this balancing point is dependent on a complex's molecular weight but not on its particular shape, the molecular weight can be directly calculated, as needed to determine the stoichiometry of each protein in a protein complex.

## Sets of Interacting Proteins Can Be Identified by Biochemical Methods

Because most proteins in the cell function as part of complexes with other proteins, an important way to begin to characterize the biological role of an unknown protein is to identify all of the other proteins to which it specifically binds.

One method for identifying proteins that bind to one another tightly is *co-immunoprecipitation*. In this case, an antibody recognizes a specific target protein; reagents that bind to the antibody and are coupled to a solid matrix then drag the complex out of solution to the bottom of a test tube. If the original target protein is associated tightly enough with another protein when it is captured by the antibody, the partner precipitates as well. This method is useful for identifying proteins that are part of a complex inside cells, including those that interact only transiently—for example, when extracellular signal molecules stimulate cells (discussed in Chapter 15). Another method frequently used to identify a protein's binding partners is protein affinity chromatography (see Figure 8–13C). To employ this technique to capture interacting proteins, a target protein is attached to polymer beads that are packed into a column. When the proteins in a cell extract are washed through this column, those proteins that interact with the target protein are retained by the affinity matrix. These proteins can then be eluted and their identity determined by mass spectrometry.

In addition to capturing protein complexes on columns or in test tubes, researchers are developing high-density protein arrays to investigate protein interactions. These arrays, which contain thousands of different proteins or antibodies spotted onto glass slides or immobilized in tiny wells, allow one to examine the biochemical activities and binding profiles of a large number of proteins at once. For example, if one incubates a fluorescently labeled protein with arrays containing thousands of immobilized proteins, the spots that remain fluorescent after extensive washing each contain a protein to which the labeled protein specifically binds.

## Protein–Protein Interactions Can Also Be Identified by a Two-Hybrid Technique in Yeast

Thus far, we have emphasized biochemical approaches to the study of protein–protein interactions. However, a particularly powerful strategy, called the **two-hybrid system**, relies on exploiting the cell's own mechanisms to reveal protein–protein interactions.

The technique takes advantage of the modular nature of gene activator proteins (see Figure 7–45). These proteins both bind to specific DNA sequences and activate gene transcription, and these activities are often performed by two separate protein domains. Using recombinant DNA techniques, two such protein domains are used to create separate "bait" and "prey" fusion proteins. To create the "bait" fusion protein, the DNA sequence that codes for a target protein is fused with DNA that encodes the DNA-binding domain of a gene activator protein. When this construct is introduced into yeast, the cells produce the fusion protein, with the target protein attached to this DNA-binding domain (**Figure 8–24**). This fusion protein binds to the regulatory region of a reporter gene, where it serves as "bait" to fish for proteins that interact with the target protein. To search for potential binding partners (potential prey for the bait), the candidate proteins also have to be constructed as fusion proteins: DNA encoding the activation domain of a gene activator protein is fused to a large

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 181 of 240

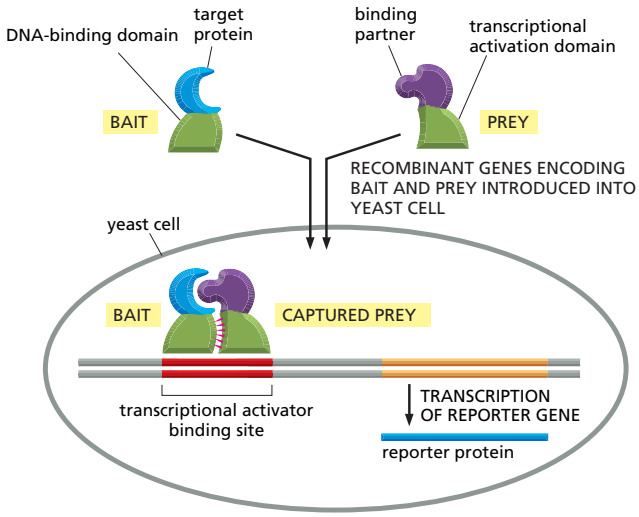



**Figure 8–24 The yeast two-hybrid system for detecting protein–protein interactions.** The target protein is fused to a DNA-binding domain that directs the fusion protein to the regulatory region of a reporter gene as "bait." When this target protein binds to another specially designed protein in the cell nucleus ("prey"), their interaction brings together two halves of a transcriptional activator, which then switches on the expression of the reporter gene.

number of different genes. Members of this collection of genes—encoding potential "prey"—are introduced individually into yeast cells containing the bait. If the yeast cell receives a DNA clone that expresses a prey partner for the bait protein, the two halves of a transcriptional activator are united, switching on the reporter gene (see Figure 8–24).

This ingenious technique sounds complex, but the two-hybrid system is relatively simple to use in the laboratory. Although the protein–protein interactions occur in the yeast cell nucleus, proteins from every part of the cell and from any organism can be studied in this way. The two-hybrid system has been scaled up to map the interactions that occur among all of the proteins an organism produces. In this case, a set of bait and prey fusions is produced for every cell protein, and every bait/prey combination can be monitored. In this way protein interaction maps have been generated for most of the proteins in yeast, *C. elegans,* and *Drosophila*.

## Combining Data Derived from Different Techniques Produces Reliable Protein-Interaction Maps

As previously discussed in Chapter 3, extensive protein-interaction maps can be very useful for identifying the functions of proteins (see Figure 3–82). For this reason, both the two-hybrid method and the biochemical technique discussed earlier known as tap-tagging (see pp. 515–516) have been automated to determine the interactions between thousands of proteins. Unfortunately, different results are found in different experiments, and many of the interactions detected in one laboratory are not detected in another. Therefore, the most useful protein-interaction maps are those that combine data from many experiments, requiring that each interaction in the map be confirmed by more than one technique.

## Optical Methods Can Monitor Protein Interactions in Real Time

Once two proteins—or a protein and a small molecule—are known to associate, it becomes important to characterize their interaction in more detail. Proteins can associate with each other more or less permanently (like the subunits of RNA polymerase or the proteosome), or engage in transient encounters that may last only a few milliseconds (like a protein kinase and its substrate).

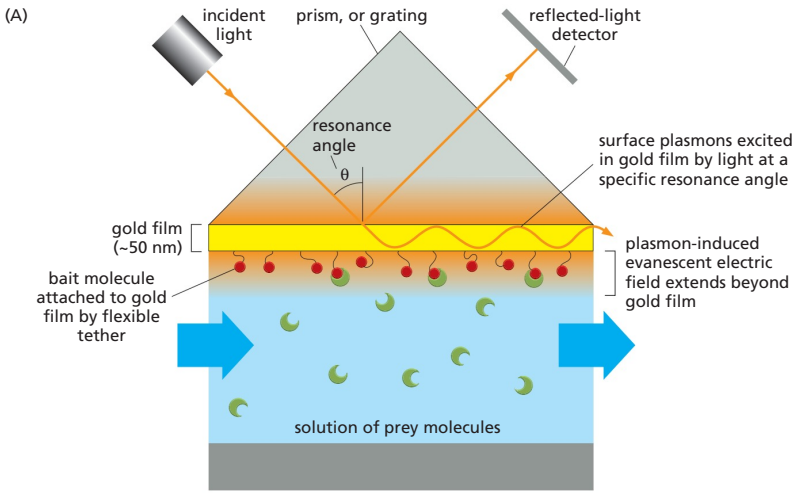
To understand how a protein functions inside a cell, we need to determine how tightly it binds to other proteins, how rapidly it dissociates from them, and how covalent modifications, small molecules, or other proteins influence these interactions. Such studies of protein dynamics often employ optical methods.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 182 of 240

Certain amino acids (for example, tryptophan) exhibit weak fluorescence that can be detected with sensitive fluorimeters. In many cases, the fluorescence intensity, or the emission spectrum of fluorescent amino acids located in a protein–protein interface, will change when the proteins associate. When this change can be detected by fluorimetry, it provides a sensitive and quantitative measure of protein binding.

A particularly useful method for monitoring the dynamics of a protein's binding to other molecules is called **surface plasmon resonance** (**SPR**). The SPR method has been used to characterize a wide variety of molecular interactions, including antibody-antigen binding, ligand-receptor coupling, and the binding of proteins to DNA, carbohydrates, small molecules, and other proteins.

SPR detects binding interactions by monitoring the reflection of a beam of light off the interface between an aqueous solution of potential binding molecules and a biosensor surface carrying an immobilized bait protein. The bait protein is attached to a very thin layer of metal that coats one side of a glass prism (**Figure 8–25**). A light beam is passed through the prism; at a certain angle, called the *resonance angle*, some of the energy from the light interacts with the cloud of electrons in the metal film, generating a plasmon—an oscillation of the electrons at right angles to the plane of the film, bouncing up and down between its upper and lower surfaces like a weight on a spring. The plasmon, in turn, generates an electrical field that extends a short distance—about the wavelength of the light—above and below the metal surface. Any change in the composition of




(B) The binding of prey molecules to bait molecules increases the refractive index of the surface layer. This alters the resonance angle for plasmon induction, which can be measured by a detector.

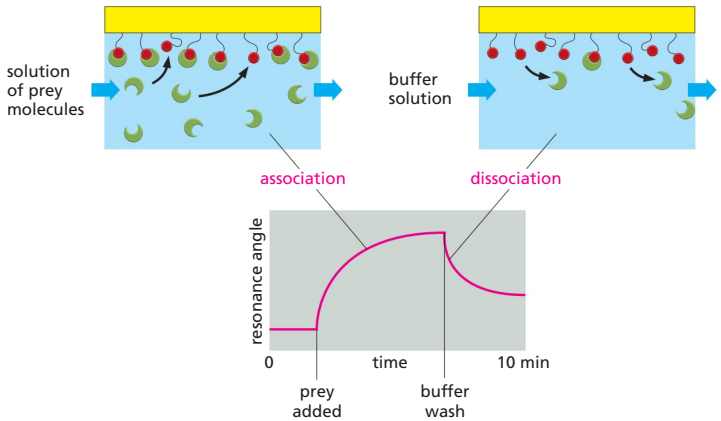



**Figure 8–25 Surface plasmon resonance.** (A) SPR can detect binding interactions by monitoring the reflection of a beam of light off the interface between an aqueous solution of potential binding molecules *(green)* and a biosensor surface coated with an immobilized bait protein *(red)*. (B) A solution of prey proteins is allowed to flow past the immobilized bait protein. Binding of prey molecules to the bait protein produces a measurable change in the resonance angle, as does their dissociation when a buffer solution washes them off. These changes, monitored in real time, reflect the association and dissociation of the molecular complexes.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 183 of 240

the environment within the range of the electrical field will cause a measurable change in the resonance angle.

To measure binding, a solution containing proteins (or other molecules) that might interact with the immobilized bait protein is allowed to flow past the biosensor surface. Proteins binding to the bait change the composition of the molecular complexes on the metal surface, causing a change in the resonance angle (see Figure 8–25). The changes in the resonance angle are monitored in real time and reflect the kinetics of the association—or dissociation—of molecules with the bait protein. The association rate ($k_{on}$) is measured as the molecules interact, and the dissociation rate ($k_{off}$) is determined as buffer washes the bound molecules from the sensor surface. A binding constant ($K$) is calculated by dividing $k_{off}$ by $k_{on}$. In addition to determining the kinetics, SPR can be used to determine the number of molecules that are bound in each complex: the magnitude of the SPR signal change is proportional to the mass of the immobilized complex.

The SPR method is particularly useful because it requires only small amounts of the protein, the protein does not have to be labeled in any way, and the interactions of the protein with other molecules can be monitored in real time.

A third optical method for probing protein interactions uses *green fluorescent protein* (discussed in detail below) and its derivatives of different colors. In this application, two proteins of interest are each labeled with a different fluorochrome, such that the emission spectrum of one fluorochrome overlaps the absorption spectrum of the second fluorochrome. If the two proteins—and their attached fluorochromes—come very close to each other (within about 1–10 nm), the energy of the absorbed light is transferred from one fluorochrome to the other. The energy transfer, called **fluorescence resonance energy transfer** (**FRET**), is determined by illuminating the first fluorochrome and measuring emission from the second (**Figure 8–26**). This technique is especially powerful because, when combined with fluorescence microscopy, it can be used to characterize protein-protein interactions at specific locations inside living cells.

## Some Techniques Can Monitor Single Molecules

The biochemical methods described so far in this chapter are used to study large populations of molecules, a limitation that reflects the small size of typical biological molecules relative to the sensitivity of the methods to detect them. However, the recent development of highly sensitive and precise measurement methods has created a new branch of biophysics—the study of single molecules. Single-molecule studies are particularly important in cell biology because many processes rely on the activities of only a few critical molecules in the cell.

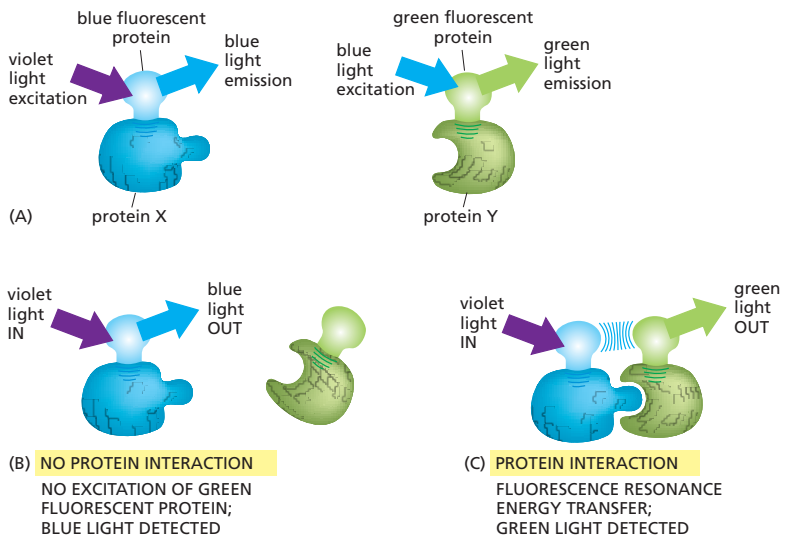



**Figure 8–26 Fluorescence resonance energy transfer (FRET).** To determine whether (and when) two proteins interact inside a cell, the proteins are first produced as fusion proteins attached to different color variants of green fluorescent protein (GFP). (A) In this example, protein X is coupled to a blue fluorescent protein, which is excited by violet light (370–440 nm) and emits blue light (440–480 nm); protein Y is coupled to a green fluorescent protein, which is excited by blue light and emits green light (510 nm). (B) If protein X and Y do not interact, illuminating the sample with violet light yields fluorescence from the blue fluorescent protein only. (C) When protein X and protein Y interact, FRET can now occur. Illuminating the sample with violet light excites the blue fluorescent protein, whose emission in turn excites the green fluorescent protein, resulting in an emission of green light. The fluorochromes must be quite close together—within about 1–10 nm of one another—for FRET to occur. Because not every molecule of protein X and protein Y is bound at all times, some blue light may still be detected. But as the two proteins begin to interact, emission from the donor GFP falls as the emission from the acceptor GFP rises.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 184 of 240

The first example of a technique for studying the function of single protein molecules was the use of a patch electrode to measure current flow through single ion channels (see Figure 11–33). Another approach is to attach the protein to a larger structure, such as a polystyrene bead, which can then be observed by conventional microscopy. This strategy has been particularly useful in measuring the movements of motor proteins. For example, molecules of the motor protein kinesin (discussed in Chapter 16) can be attached to a bead, and by observing the kinesin-attached bead moving along a microtubule, the step size of the motor (that is, the distance moved for each ATP molecule hydrolyzed) can be measured. As we will see in Chapter 9, optical microscopes have a limited resolution due to the diffraction of light, but computational and optical methods can be used to determine the position of a bead to a much finer precision than the resolution limit of the microscope. Using such techniques, extremely small movements—on the order of nanometers—can easily be detected and quantified.

Another advantage of attaching molecules to large beads is that these beads can serve as "handles" by which the molecules can be manipulated. This allows forces to be applied to the molecules, and their response observed. For example, the speed or step size of a motor can be measured as a function of the force it is pulling against. As discussed in the next chapter, a focused laser beam can be used as "optical tweezers" to generate a mechanical force on a bead, allowing motor proteins to be studied under an applied force (see Figure 9–35). <CGCG> Beads can also be manipulated using magnetic fields, a technology known as "magnetic tweezers." If multiple beads are present in a magnetic field, they will all experience the same force, potentially allowing large numbers of beads to be manipulated in parallel in a single experiment.

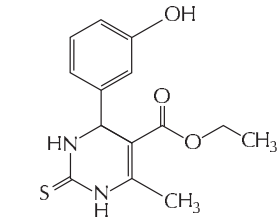
While beads can be used as markers to track protein movements, it is clearly preferable to be able to visualize the proteins themselves. In the next chapter, we shall see that recent refinements in microscopy have now made this possible.

## Protein Function Can Be Selectively Disrupted With Small Molecules

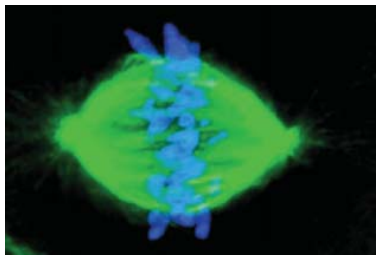
Chemical inhibitors have contributed to the development of cell biology. For example, the microtubule inhibitor colchicine is routinely used to test whether microtubules are required for a given biological process; it also led to the first purification of tubulin several decades ago. In the past, these small molecules were usually natural products; that is, they were synthesized by living creatures. Although, as a whole, natural products have been extraordinarily useful in science and medicine (see, for example, Table 6–4, p. 385), they acted on a limited number of biological processes. However, the recent development of methods to synthesize hundreds of thousands of small molecules and to carry out large-scale automated screens holds the promise of identifying chemical inhibitors for virtually any biological process. In such approaches, large collections of small chemical compounds are simultaneously tested, either on living cells or in cell-free assays. Once an inhibitor is identified, it can be used as a probe to identify, through affinity chromatography (see Figure 8–13C) or other means, the protein to which the inhibitor binds. This general strategy, often called **chemical biology**, has successfully identified inhibitors of many proteins that carry out key processes in cell biology. The kinesin protein that functions in mitosis, for example, was identified by this method (**Figure 8–27**). Chemical inhibitors give the cell biologist great control over the timing of inhibition, as drugs can be rapidly added to or removed from cells, allowing protein function to be switched on or off quickly.

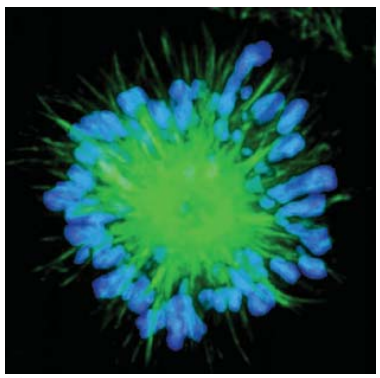
## Protein Structure Can Be Determined Using X-Ray Diffraction

The main technique that has been used to discover the three-dimensional structure of molecules, including proteins, at atomic resolution is **x-ray crystallography**. X-rays, like light, are a form of electromagnetic radiation, but they have a much shorter wavelength, typically around 0.1 nm (the diameter of a hydrogen



(A) monastrol

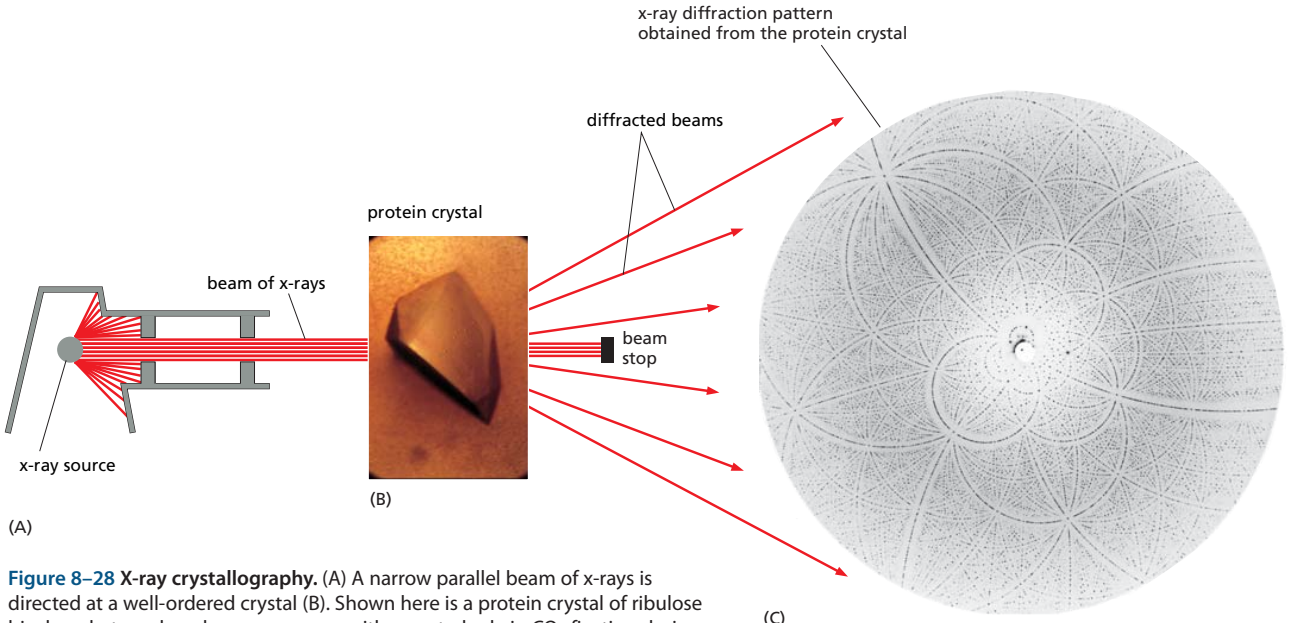

(B)


(C)

5 μm

**Figure 8–27 Small-molecule inhibitors for manipulating living cells.** (A) Chemical structure of monastrol, a kinesin inhibitor identified in a large-scale screen for small molecules that disrupt mitosis. (B) Normal mitotic spindle seen in an untreated cell. The microtubules are stained *green* and chromosomes *blue*. (C) Monopolar spindle that forms in cells treated with monastrol. (B and C, from T.U. Mayer et al., *Science* 286:971–974, 1999. With permission from AAAS.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 185 of 240



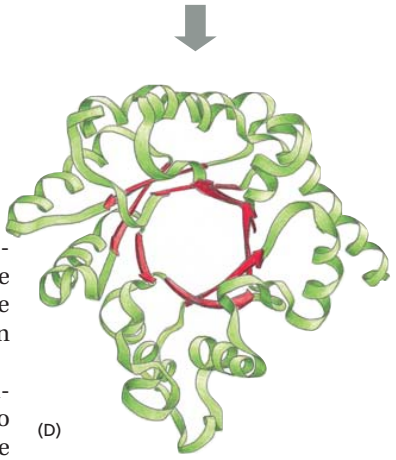




**Figure 8–28 X-ray crystallography.** (A) A narrow parallel beam of x-rays is directed at a well-ordered crystal (B). Shown here is a protein crystal of ribulose bisphosphate carboxylase, an enzyme with a central role in $CO_2$ fixation during photosynthesis. The atoms in the crystal scatter some of the beam, and the scattered waves reinforce one another at certain points and appear as a pattern of diffraction spots (C). This diffraction pattern, together with the amino acid sequence of the protein, can be used to produce an atomic model (D). The complete atomic model is hard to interpret, but this simplified version, derived from the x-ray diffraction data, shows the protein's structural features clearly (α helices, *green*; β strands, *red*). The components pictured in A to D are not shown to scale. (B, courtesy of C. Branden; C, courtesy of J. Hajdu and I. Andersson; D, adapted from original provided by B. Furugren.)

atom). If a narrow parallel beam of x-rays is directed at a sample of a pure protein, most of the x-rays pass straight through it. A small fraction, however, are scattered by the atoms in the sample. If the sample is a well-ordered crystal, the scattered waves reinforce one another at certain points and appear as diffraction spots when recorded by a suitable detector (**Figure 8–28**).

The position and intensity of each spot in the x-ray diffraction pattern contain information about the locations of the atoms in the crystal that gave rise to it. Deducing the three-dimensional structure of a large molecule from the diffraction pattern of its crystal is a complex task and was not achieved for a protein molecule until 1960. But in recent years x-ray diffraction analysis has become increasingly automated, and now the slowest step is likely to be the generation of suitable protein crystals. This step requires large amounts of very pure protein and often involves years of trial and error to discover the proper crystallization conditions; the pace has greatly accelerated with the use of recombinant DNA techniques to produce pure proteins and robotic techniques to test large numbers of crystallization conditions.

Analysis of the resulting diffraction pattern produces a complex three-dimensional electron-density map. Interpreting this map—translating its contours into a three-dimensional structure—is a complicated procedure that requires knowledge of the amino acid sequence of the protein. Largely by trial and error, the sequence and the electron-density map are correlated by computer to give the best possible fit. The reliability of the final atomic model depends on the resolution of the original crystallographic data: 0.5 nm resolution might produce a low-resolution map of the polypeptide backbone, whereas a resolution of 0.15 nm allows all of the non-hydrogen atoms in the molecule to be reliably positioned.

A complete atomic model is often too complex to appreciate directly, but simplified versions that show a protein's essential structural features can be readily derived from it (see Panel 3–2, pp. 132–133). The three-dimensional

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 186 of 240

structures of about 20,000 different proteins have now been determined by x-ray crystallography or by NMR spectroscopy (see below)—enough to begin to see families of common structures emerging. These structures or protein folds often seem to be more conserved in evolution than are the amino acid sequences that form them (see Figure 3–13).

X-ray crystallographic techniques can also be applied to the study of macromolecular complexes. In a recent triumph, the method was used to determine the structure of the ribosome, a large and complex machine made of several RNAs and more than 50 proteins (see Figure 6–64). The determination required the use of a synchrotron, a radiation source that generates x-rays with the intensity needed to analyze the crystals of such large macromolecular complexes. <GGCC>

## NMR Can Be Used to Determine Protein Structure in Solution

**Nuclear magnetic resonance** (**NMR**) **spectroscopy** has been widely used for many years to analyze the structure of small molecules. This technique is now also increasingly applied to the study of small proteins or protein domains. Unlike x-ray crystallography, NMR does not depend on having a crystalline sample. It simply requires a small volume of concentrated protein solution that is placed in a strong magnetic field; indeed, it is the main technique that yields detailed evidence about the three-dimensional structure of molecules in solution.

Certain atomic nuclei, particularly hydrogen nuclei, have a magnetic moment or spin: that is, they have an intrinsic magnetization, like a bar magnet. The spin aligns along the strong magnetic field, but it can be changed to a misaligned, excited state in response to applied radiofrequency (RF) pulses of electromagnetic radiation. When the excited hydrogen nuclei return to their aligned state, they emit RF radiation, which can be measured and displayed as a spectrum. The nature of the emitted radiation depends on the environment of each hydrogen nucleus, and if one nucleus is excited, it influences the absorption and emission of radiation by other nuclei that lie close to it. It is consequently possible, by an ingenious elaboration of the basic NMR technique known as two-dimensional NMR, to distinguish the signals from hydrogen nuclei in different amino acid residues, and to identify and measure the small shifts in these signals that occur when these hydrogen nuclei lie close enough together to interact. Because the size of such a shift reveals the distance between the interacting pair of hydrogen atoms, NMR can provide information about the distances between the parts of the protein molecule. By combining this information with a knowledge of the amino acid sequence, it is possible in principle to compute the three-dimensional structure of the protein (**Figure 8–29**).

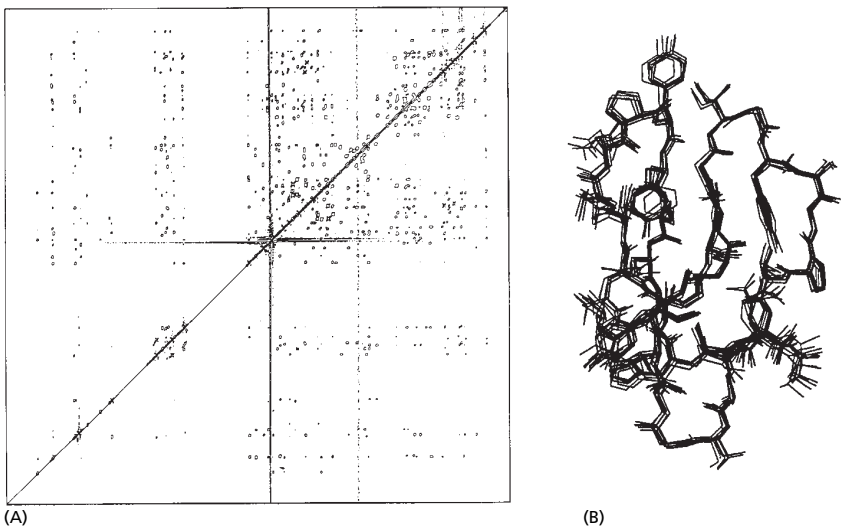



**Figure 8–29 NMR spectroscopy.** (A) An example of the data from an NMR machine. This two-dimensional NMR spectrum is derived from the C-terminal domain of the enzyme cellulase. The spots represent interactions between hydrogen atoms that are near neighbors in the protein and hence reflect the distance that separates them. Complex computing methods, in conjunction with the known amino acid sequence, enable possible compatible structures to be derived. (B) Ten structures of the enzyme, which all satisfy the distance constraints equally well, are shown superimposed on one another, giving a good indication of the probable three-dimensional structure. (Courtesy of P. Kraulis.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 187 of 240

For technical reasons the structure of small proteins of about 20,000 daltons or less can be most readily determined by NMR spectroscopy. Resolution decreases as the size of a macromolecule increases. But recent technical advances have now pushed the limit to about 100,000 daltons, thereby making the majority of proteins accessible for structural analysis by NMR.

Because NMR studies are performed in solution, this method also offers a convenient means of monitoring changes in protein structure, for example during protein folding or when the protein binds to another molecule. NMR is also used widely to investigate molecules other than proteins and is valuable, for example, as a method to determine the three-dimensional structures of RNA molecules and the complex carbohydrate side chains of glycoproteins.

Some landmarks in the development of x-ray crystallography and NMR are listed in **Table 8–2**.

## Protein Sequence and Structure Provide Clues About Protein Function

Having discussed methods for purifying and analyzing proteins, we now turn to a common situation in cell and molecular biology: an investigator has identified a gene important for a biological process but has no direct knowledge of the biochemical properties of its protein product.

Thanks to the proliferation of protein and nucleic acid sequences that are catalogued in genome databases, the function of a gene—and its encoded protein—can often be predicted by simply comparing its sequence with those of previously characterized genes (see Figure 3–14). Because amino acid sequence

**Table 8–2 Landmarks in the Development of X-ray Crystallography and NMR and Their Application to Biological Molecules**

| | |
|---|---|
| 1864 | **Hoppe-Seyler** crystallizes, and names, the protein hemoglobin. |
| 1895 | **Röntgen** observes that a new form of penetrating radiation, which he names x-rays, is produced when cathode rays (electrons) hit a metal target. |
| 1912 | **Von Laue** obtains the first x-ray diffraction patterns by passing x-rays through a crystal of zinc sulfide.<br>**W.L. Bragg** proposes a simple relationship between an x-ray diffraction pattern and the arrangement of atoms in a crystal that produce the pattern. |
| 1926 | **Summer** obtains crystals of the enzyme urease from extracts of jack beans and demonstrates that proteins possess catalytic activity. |
| 1931 | **Pauling** publishes his first essays on 'The Nature of the Chemical Bond,' detailing the rules of covalent bonding. |
| 1934 | **Bernal and Crowfoot** present the first detailed x-ray diffraction patterns of a protein obtained from crystals of the enzyme pepsin. |
| 1935 | **Patterson** develops an analytical method for determining interatomic spacings from x-ray data. |
| 1941 | **Astbury** obtains the first x-ray diffraction pattern of DNA. |
| 1946 | **Block** and **Purcell** describe NMR. |
| 1951 | **Pauling and Corey** propose the structure of a helical conformation of a chain of L-amino acids—the α helix—and the structure of the β sheet, both of which were later found in many proteins. |
| 1953 | **Watson and Crick** propose the double-helix model of DNA, based on x-ray diffraction patterns obtained by **Franklin and Wilkins**. |
| 1954 | **Perutz** and colleagues develop heavy-atom methods to solve the phase problem in protein crystallography. |
| 1960 | **Kendrew** describes the first detailed structure of a protein (sperm whale myoglobin) to a resolution of 0.2 nm, and **Perutz** presents a lower-resolution structure of the larger protein hemoglobin. |
| 1966 | **Phillips** describes the structure of lysozyme, the first enzyme to have its structure analyzed in detail. |
| 1971 | **Jeener** proposes the use of two-dimensional NMR, and **Wuthrich** and colleagues first use the method to solve a protein structure in the early 1980s. |
| 1976 | **Kim and Rich** and **Klug** and colleagues describe the detailed three-dimensional structure of tRNA determined by x-ray diffraction. |
| 1977–1978 | **Holmes and Klug** determine the structure of tobacco mosaic virus (TMV), and **Harrison** and **Rossman** determine the structure of two small spherical viruses. |
| 1985 | **Michel**, **Deisenhofer** and colleagues determine the first structure of a transmembrane protein (a bacterial reaction center) by x-ray crystallography. **Henderson** and colleagues obtain the structure of bacteriorhodopsin, a transmembrane protein, by high-resolution electron-microscopy methods between 1975 and 1990. |

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 188 of 240

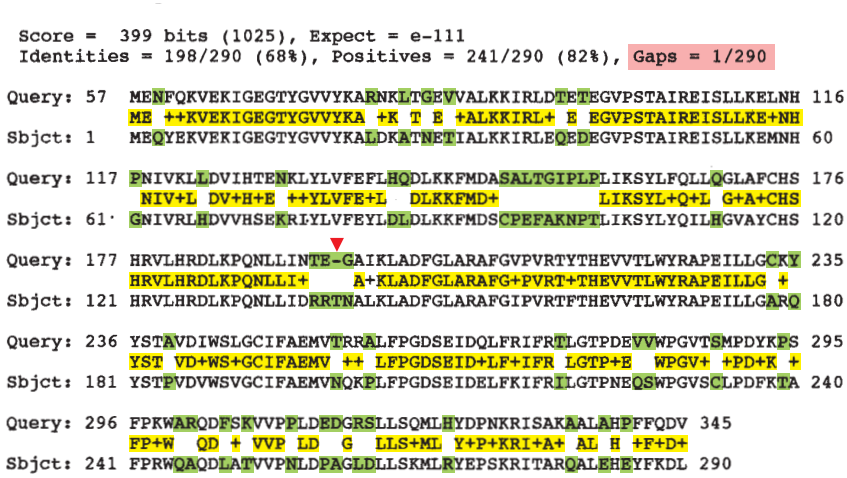



**Figure 8–30 Results of a BLAST search.** Sequence databases can be searched to find similar amino acid or nucleic acid sequences. Here, a search for proteins similar to the human cell-cycle regulatory protein Cdc2 *(Query)* locates maize Cdc2 *(Sbjct)*, which is 68% identical (and 82% similar) to human Cdc2 in its amino acid sequence. The alignment begins at residue 57 of the Query protein, suggesting that the human protein has an N-terminal region that is absent from the maize protein. The *green* blocks indicate differences in sequence, and the *yellow* bar summarizes the similarities: when the two amino acid sequences are identical, the residue is shown; conservative amino acid substitutions are indicated by a plus sign (+). Only one small gap has been introduced—indicated by the *red arrow* at position 194 in the Query sequence—to align the two sequences maximally. The alignment score *(Score)*, which is expressed in two different types of units, takes into account penalties for substitutions and gaps; the higher the alignment score, the better the match. The significance of the alignment is reflected in the *Expectation* (E) value, which specifies how often a match this good would be expected to occur by chance. The lower the E value, the more significant the match; the extremely low value here ($e^{-111}$) indicates certain significance. E values much higher than 0.1 are unlikely to reflect true relatedness. For example, an E value of 0.1 means there is a 1 in 10 likelihood that such a match would arise solely by chance.

determines protein structure, and structure dictates biochemical function, proteins that share a similar amino acid sequence usually have the same structure and usually perform similar biochemical functions, even when they are found in distantly related organisms. In modern cell biology, the study of a newly discovered protein usually begins with a search for previously characterized proteins that are similar in their amino acid sequences.

Searching a collection of known sequences for homologous genes or proteins is typically done over the World Wide Web, and it simply involves selecting a database and entering the desired sequence. A sequence alignment program—the most popular are BLAST and FASTA—scans the database for similar sequences by sliding the submitted sequence along the archived sequences until a cluster of residues falls into full or partial alignment (**Figure 8–30**). The results of even a complex search—which can be performed on either a nucleotide or an amino acid sequence—are returned within minutes. Such comparisons can predict the functions of individual proteins, families of proteins, or even most of the protein complement of a newly sequenced organism.

As was explained in Chapter 3, many proteins that adopt the same conformation and have related functions are too distantly related to be identified as clearly homologous from a comparison of their amino acid sequences alone (see Figure 3–13). Thus, an ability to reliably predict the three dimensional structure of a protein from its amino acid sequence would improve our ability to infer protein function from the sequence information in genomic databases. In recent years, major progress has been made in predicting the precise structure of a protein. These predictions are based, in part, on our knowledge of tens of thousands of protein structures that have already been determined by x-ray crystallography and NMR spectroscopy and, in part, on computations using our knowledge of the physical forces acting on the atoms. However, it remains a substantial and important challenge to predict the structures of proteins that are large or have multiple domains, or to predict structures at the very high levels of resolution needed to assist in computer-based drug discovery.

While finding homologous sequences and structures for a new protein will provide many clues about its function, it is usually necessary to test these insights through direct experimentation. However, the clues generated from sequence comparisons typically point the investigator in the correct experimental direction, and their use has therefore become one of the most important strategies in modern cell biology.

## Summary

*Most proteins function in concert with other proteins, and many methods exist for identifying and studying protein–protein interactions. Small-molecule inhibitors allow the functions of proteins they act upon to be studied in living cells. Because proteins with similar structures often have similar functions, the biochemical activity of a*

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 189 of 240

*protein can often be predicted by searching databases for previously characterized proteins that are similar in their amino acid sequences.*

## ANALYZING AND MANIPULATING DNA

Until the early 1970s, DNA was the most difficult biological molecule for the biochemist to analyze. Enormously long and chemically monotonous, the string of nucleotides that forms the genetic material of an organism could be examined only indirectly, by protein or RNA sequencing or by genetic analysis. Today, the situation has changed entirely. From being the most difficult macromolecule of the cell to analyze, DNA has become the easiest. It is now possible to isolate a specific region of almost any genome, to produce a virtually unlimited number of copies of it, and to determine the sequence of its nucleotides in a few hours. At the height of the Human Genome Project, large facilities with automated machines were generating DNA sequences at the rate of 1000 nucleotides per second, around the clock. By related techniques, an isolated gene can be altered (engineered) at will and transferred back into the germ line of an animal or plant, so as to become a functional and heritable part of the organism's genome.

These technical breakthroughs in **genetic engineering**—the ability to manipulate DNA with precision in a test tube or an organism—have had a dramatic impact on all aspects of cell biology by facilitating the study of cells and their macromolecules in previously unimagined ways. **Recombinant DNA technology** comprises a mixture of techniques, some newly developed and some borrowed from other fields such as microbial genetics (**Table 8–3**). Central to the technology are the following key techniques:

1. Cleavage of DNA at specific sites by restriction nucleases, which greatly facilitates the isolation and manipulation of individual genes.
2. DNA ligation, which makes it possible to design and construct DNA molecules that are not found in nature.
3. DNA cloning through the use of either cloning vectors or the polymerase chain reaction, in which a portion of DNA is repeatedly copied to generate many billions of identical molecules.
4. Nucleic acid hybridization, which makes it possible to find a specific sequence of DNA or RNA with great accuracy and sensitivity on the basis of its ability to selectively bind a complementary nucleic acid sequence.
5. Rapid determination of the sequence of nucleotides of any DNA (even entire genomes), making it possible to identify genes and to deduce the amino acid sequence of the proteins they encode.
6. Simultaneous monitoring of the level of mRNA produced by every gene in a cell, using nucleic acid microarrays, in which tens of thousands of hybridization reactions take place simultaneously.

In this section, we describe each of these basic techniques, which together have revolutionized the study of cell biology.

### Restriction Nucleases Cut Large DNA Molecules into Fragments

Unlike a protein, a gene does not exist as a discrete entity in cells, but rather as a small region of a much longer DNA molecule. Although the DNA molecules in a cell can be randomly broken into small pieces by mechanical force, a fragment containing a single gene in a mammalian genome would still be only one among a hundred thousand or more DNA fragments, indistinguishable in their average size. How could such a gene be purified? Because all DNA molecules consist of an approximately equal mixture of the same four nucleotides, they cannot be readily separated, as proteins can, on the basis of their different charges and binding properties.

The solution to all of these problems began to emerge with the discovery of **restriction nucleases**. These enzymes, which can be purified from bacteria, cut the DNA double helix at specific sites defined by the local nucleotide sequence, thereby cleaving a long double-stranded DNA molecule into fragments of

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 190 of 240

**Table 8–3 Some Major Steps in the Development of Recombinant DNA and Transgenic Technology**

| | |
|---|---|
| 1869 | **Miescher** first isolates DNA from white blood cells harvested from pus-soaked bandages obtained from a nearby hospital. |
| 1944 | **Avery** provides evidence that DNA, rather than protein, carries the genetic information during bacterial transformation. |
| 1953 | **Watson and Crick** propose the double-helix model for DNA structure based on x-ray results of **Franklin and Wilkins**. |
| 1955 | **Kornberg** discovers DNA polymerase, the enzyme now used to produce labeled DNA probes. |
| 1961 | **Marmur and Doty** discover DNA renaturation, establishing the specificity and feasibility of nucleic acid hydridization reactions. |
| 1962 | **Arber** provides the first evidence for the existence of DNA restriction nucleases, leading to their purification and use in DNA sequence characterization by **Nathans and H. Smith**. |
| 1966 | **Nirenberg**, **Ochoa**, and **Khorana** elucidate the genetic code. |
| 1967 | **Gellert** discovers DNA ligase, the enzyme used to join DNA fragments together. |
| 1972–1973 | DNA cloning techniques are developed by the laboratories of **Boyer**, **Cohen**, **Berg**, and their colleagues at Stanford University and the University of California at San Francisco. |
| 1975 | **Southern** develops gel-transfer hybridization for the detection of specific DNA sequences. |
| 1975–1977 | **Sanger and Barrell** and **Maxam and Gilbert** develop rapid DNA-sequencing methods. |
| 1981–1982 | **Palmiter and Brinster** produce transgenic mice; **Spradling and Rubin** produce transgenic fruit flies. |
| 1982 | **GenBank**, NIH's public genetic sequence database, is established at Los Alamos National Laboratory. |
| 1985 | **Mullis** and co-workers invent the polymerase chain reaction (PCR). |
| 1987 | **Capecchi** and **Smithies** introduce methods for performing targeted gene replacement in mouse embryonic stem cells. |
| 1989 | **Fields and Song** develop the yeast two-hybrid system for identifying and studying protein interactions. |
| 1989 | **Olson** and colleagues describe sequence-tagged sites, unique stretches of DNA that are used to make physical maps of human chromosomes. |
| 1990 | **Lipman** and colleagues release BLAST, an algorithm used to search for homology between DNA and protein sequences. |
| 1990 | **Simon** and colleagues study how to efficiently use bacterial artificial chromosomes, BACs, to carry large pieces of cloned human DNA for sequencing. |
| 1991 | **Hood and Hunkapillar** introduce new automated DNA sequence technology. |
| 1995 | **Venter** and colleagues sequence the first complete genome, that of the bacterium *Haemophilus influenzae*. |
| 1996 | **Goffeau** and an international consortium of researchers announce the completion of the first genome sequence of a eucaryote, the yeast *Saccharomyces cerevisiae*. |
| 1996–1997 | **Lockhart** and colleagues and **Brown and DeRisi** produce DNA microarrays, which allow the simultaneous monitoring of thousands of genes. |
| 1998 | **Sulston and Waterston** and colleagues produce the first complete sequence of a multicellular organism, the nematode worm *Caenorhabditis elegans*. |
| 2001 | Consortia of researchers announce the completion of the draft human genome sequence. |
| 2004 | Publication of the "finished" human genome sequence. |

strictly defined sizes. Different restriction nucleases have different sequence specificities, and it is relatively simple to find an enzyme that can create a DNA fragment that includes a particular gene. The size of the DNA fragment can then be used as a basis for partial purification of the gene from a mixture.

Different species of bacteria make different restriction nucleases, which protect them from viruses by degrading incoming viral DNA. Each bacterial nuclease recognizes a specific sequence of four to eight nucleotides in DNA. These sequences, where they occur in the genome of the bacterium itself, are protected from cleavage by methylation at an A or a C nucleotide; the sequences in foreign DNA are generally not methylated and so are cleaved by the restriction nucleases. Large numbers of restriction nucleases have been purified from various species of bacteria; several hundred, most of which recognize different nucleotide sequences, are now available commercially.

Some restriction nucleases produce staggered cuts, which leave short single-stranded tails at the two ends of each fragment (**Figure 8–31**). Ends of this type are known as *cohesive ends,* as each tail can form complementary base pairs with the tail at any other end produced by the same enzyme (**Figure 8–32**). The cohesive ends generated by restriction enzymes allow any two DNA fragments to be easily joined together, as long as the fragments were generated with the same restriction nuclease (or with another nuclease that produces the same cohesive ends). DNA molecules produced by splicing together two or more DNA fragments are called **recombinant DNA** molecules.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 191 of 240

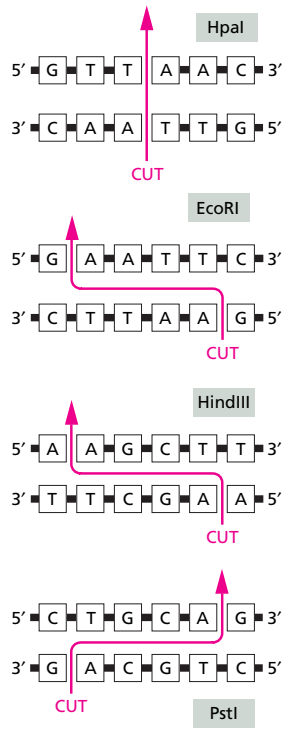



**Figure 8–31 The DNA nucleotide sequences recognized by four widely used restriction nucleases.** As in the examples shown, such sequences are often six base pairs long and "palindromic" (that is, the nucleotide sequence is the same if the helix is turned by 180 degrees around the center of the short region of helix that is recognized). The enzymes cut the two strands of DNA at or near the recognition sequence. For the genes encoding some enzymes, such as HpaI, the cleavage leaves blunt ends; for others, such as EcoRI, HindIII, and PstI, the cleavage is staggered and creates cohesive ends. Restriction nucleases are obtained from various species of bacteria: HpaI is from *Haemophilus parainfluenzae,* EcoRI is from *Escherichia coli,* HindIII is from *Haemophilus influenzae,* and PstI is from *Providencia stuartii.*

## Gel Electrophoresis Separates DNA Molecules of Different Sizes

The same types of gel electrophoresis methods that have proved so useful in the analysis of proteins can determine the length and purity of DNA molecules. The procedure is actually simpler than for proteins: because each nucleotide in a nucleic acid molecule already carries a single negative charge (on the phosphate group), there is no need to add the negatively charged detergent SDS that is required to make protein molecules move uniformly toward the positive electrode. For DNA fragments less than 500 nucleotides long, specially designed polyacrylamide gels allow the separation of molecules that differ in length by as little as a single nucleotide (**Figure 8–33**A). The pores in polyacrylamide gels, however, are too small to permit very large DNA molecules to pass; to separate these by size, the much more porous gels formed by dilute solutions of agarose (a polysaccharide isolated from seaweed) are used (Figure 8–33B). These DNA separation methods are widely used for both analytical and preparative purposes.

A variation of agarose-gel electrophoresis, called *pulsed-field gel electrophoresis,* makes it possible to separate even extremely long DNA molecules. Ordinary gel electrophoresis fails to separate such molecules because the steady electric field stretches them out so that they travel end-first through the gel in snakelike configurations at a rate that is independent of their length. In pulsed-field gel electrophoresis, by contrast, the direction of the electric field changes periodically, which forces the molecules to reorient before continuing to move snakelike through the gel. This reorientation takes much more time for larger molecules, so that longer molecules move more slowly than shorter ones. As a consequence, even entire bacterial or yeast chromosomes separate into discrete bands in pulsed-field gels and so can be sorted and identified on the basis of their size (Figure 8–33C). Although a typical mammalian chromosome of $10^8$ base pairs is too large to be sorted even in this way, large segments of these chromosomes are readily separated and identified if the chromosomal DNA is first cut with a restriction nuclease selected to recognize sequences that occur only rarely (once every 10,000 or more nucleotide pairs).

The DNA bands on agarose or polyacrylamide gels are invisible unless the DNA is labeled or stained in some way. One sensitive method of staining DNA is to expose it to the dye *ethidium bromide,* which fluoresces under ultraviolet light when it is bound to DNA (see Figure 8–33B,C). An even more sensitive detection method incorporates a radioisotope into the DNA molecules before electrophoresis; $^{32}P$ is often used as it can be incorporated into DNA phosphates and emits an energetic β particle that is easily detected by autoradiography, as in Figure 8–33. (For a discussion of radioisotopes, see p. 601).

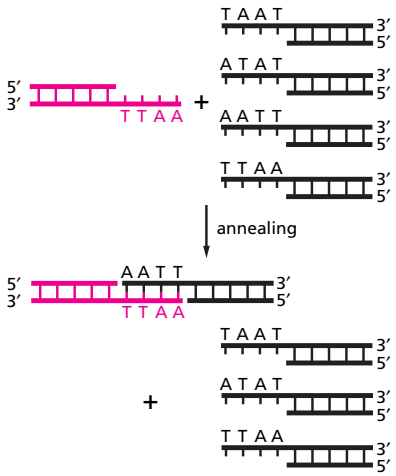



**Figure 8–32 The use of restriction nucleases to produce DNA fragments that can be easily joined together.** Fragments with the same cohesive ends can readily join by complementary base-pairing between their cohesive ends, as illustrated. The two DNA fragments that join in this example were both produced by the EcoRI restriction nuclease, whereas the three other fragments were produced by different restriction nucleases that generated different cohesive ends (see Figure 8–31). Blunt-ended fragments, like those generated by HpaI (see Figure 8–31), can be spliced together with more difficulty.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 192 of 240

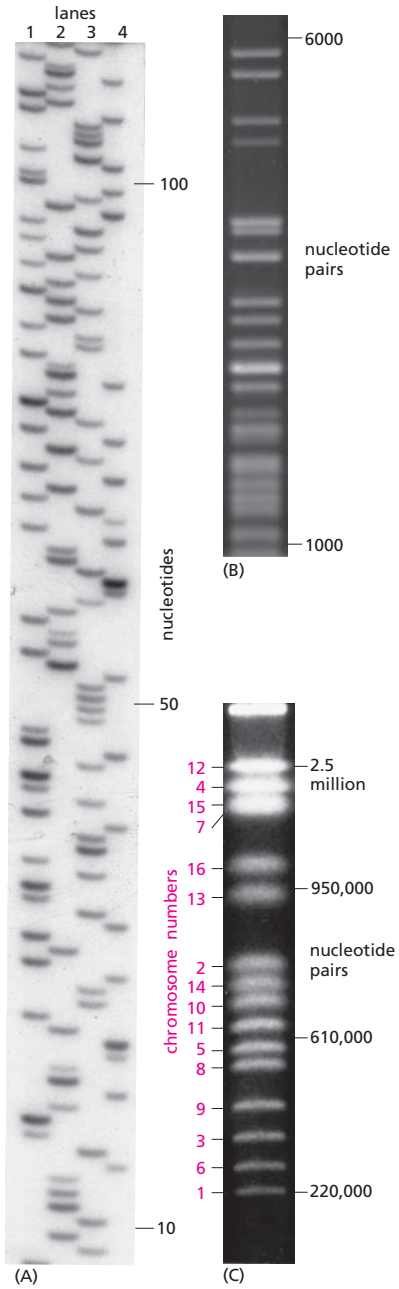



**Figure 8–33 Gel electrophoresis techniques for separating DNA molecules by size.** In the three examples shown, electrophoresis is from top to bottom, so that the largest—and thus slowest-moving—DNA molecules are near the top of the gel. (A) A polyacrylamide gel with small pores was used to fractionate single-stranded DNA. In the size range 10 to 500 nucleotides, DNA molecules that differ in size by only a single nucleotide can be separated from each other. In the example, the four lanes represent sets of DNA molecules synthesized in the course of a DNA-sequencing procedure. The DNA to be sequenced has been artificially replicated from a fixed start site up to a variable stopping point, producing a set of partial replicas of differing lengths. (Figure 8–50 explains how such sets of partial replicas are synthesized.) Lane 1 shows all the partial replicas that terminate in a G, lane 2 all those that terminate in an A, lane 3 all those that terminate in a T, and lane 4 all those that terminate in a C. Since the DNA molecules used in these reactions were radiolabeled, their positions can be determined by autoradiography, as shown.

(B) An agarose gel with medium-sized pores was used to separate double-stranded DNA molecules. This method is most useful in the size range 300 to 10,000 nucleotide pairs. These DNA molecules are fragments produced by cleaving the genome of a bacterial virus with a restriction nuclease, and they have been detected by their fluorescence when stained with the dye ethidium bromide. (C) The technique of pulsed-field agarose gel electrophoresis was used to separate 16 different yeast *(Saccharomyces cerevisiae)* chromosomes, which range in size from 220,000 to 2.5 million nucleotide pairs. The DNA was stained as in (B). DNA molecules as large as $10^7$ nucleotide pairs can be separated in this way. (A, courtesy of Leander Lauffer and Peter Walter; B, courtesy of Ken Kreuzer; C, from D. Vollrath and R.W. Davis, *Nucleic Acids Res.* 15:7865–7876, 1987. With permission from Oxford University Press.)

## Purified DNA Molecules Can Be Specifically Labeled with Radioisotopes or Chemical Markers *in vitro*

Two procedures are widely used to label isolated DNA molecules. In the first method, a DNA polymerase copies the DNA in the presence of nucleotides that are either radioactive (usually labeled with $^{32}P$) or chemically tagged (**Figure 8–34**A). In this way, "DNA probes" containing many labeled nucleotides can be produced for nucleic acid hybridization reactions (discussed below). The second procedure uses the bacteriophage enzyme polynucleotide kinase to transfer a single $^{32}P$-labeled phosphate from ATP to the 5′ end of each DNA chain (Figure 8–34B). Because only one $^{32}P$ atom is incorporated by the kinase into each DNA strand, the DNA molecules labeled in this way are often not radioactive enough to be used as DNA probes; because they are labeled at only one end, however, they have been invaluable for other applications, including DNA footprinting, as discussed in Chapter 7.

Radioactive labeling methods are being replaced by labeling with molecules that can be detected chemically or through fluorescence. To produce such nonradioactive DNA molecules, specially modified nucleotide precursors are used (Figure 8–34C). A DNA molecule made in this way is allowed to bind to its complementary DNA sequence by hybridization, as discussed in the next section, and is then detected with an antibody (or other ligand) that specifically recognizes its modified side chain (**Figure 8–35**).

## Nucleic Acid Hybridization Reactions Provide a Sensitive Way to Detect Specific Nucleotide Sequences

When an aqueous solution of DNA is heated at 100°C or exposed to a very high pH (pH ≥ 13), the complementary base pairs that normally hold the two strands of the double helix together are disrupted and the double helix rapidly dissociates into two single strands. This process, called *DNA denaturation*, was for many years thought to be irreversible. In 1961, however, it was discovered that complementary single strands of DNA readily re-form double helices by a process called **hybridization** (also called *DNA renaturation*) if they are kept for a

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 193 of 240

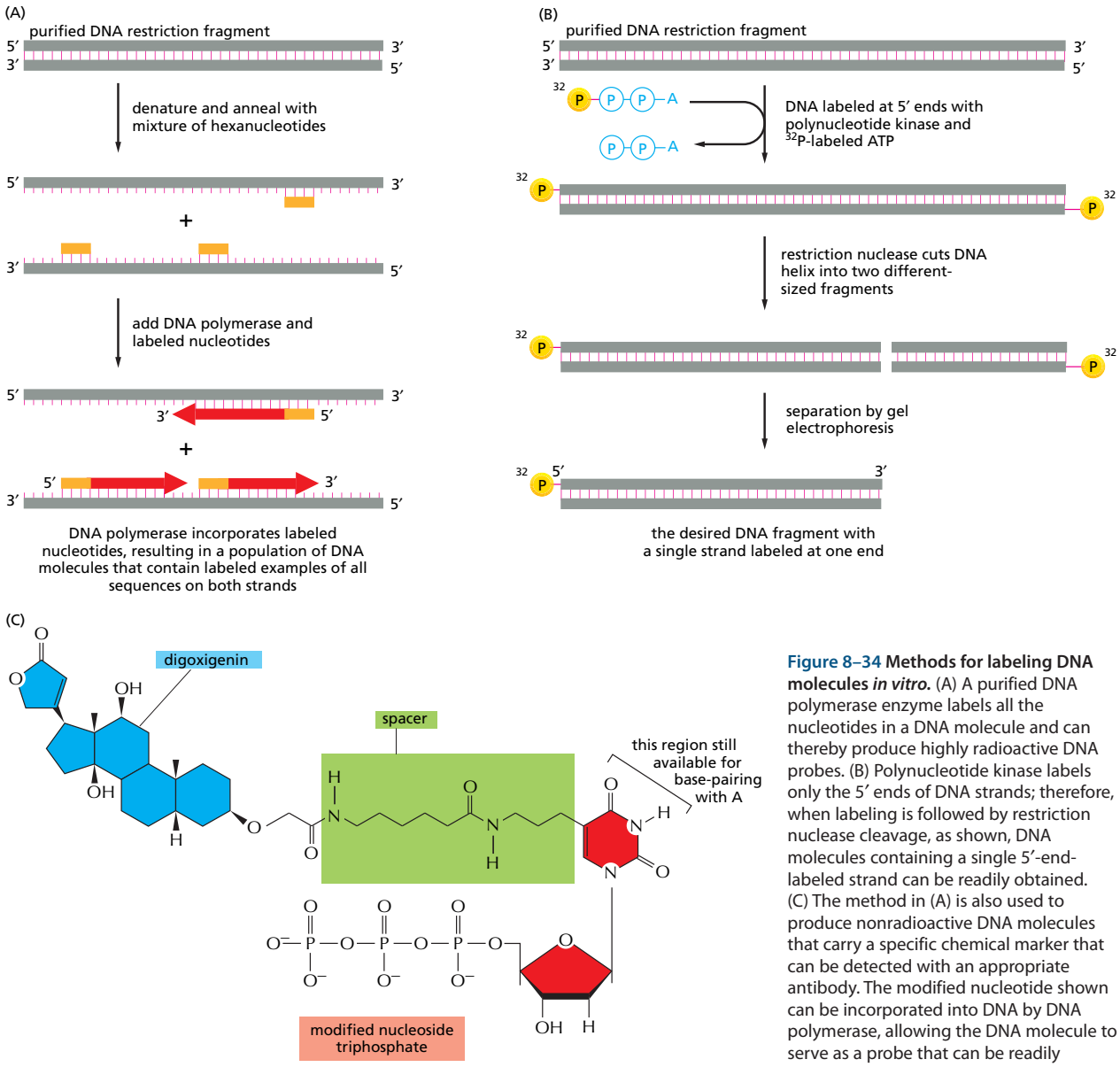



**Figure 8–34 Methods for labeling DNA molecules *in vitro*.** (A) A purified DNA polymerase enzyme labels all the nucleotides in a DNA molecule and can thereby produce highly radioactive DNA probes. (B) Polynucleotide kinase labels only the 5′ ends of DNA strands; therefore, when labeling is followed by restriction nuclease cleavage, as shown, DNA molecules containing a single 5′-end-labeled strand can be readily obtained. (C) The method in (A) is also used to produce nonradioactive DNA molecules that carry a specific chemical marker that can be detected with an appropriate antibody. The modified nucleotide shown can be incorporated into DNA by DNA polymerase, allowing the DNA molecule to serve as a probe that can be readily detected. The base on the nucleoside triphosphate shown is an analog of thymine, in which the methyl group on T has been replaced by a spacer arm linked to the plant steroid digoxigenin. An anti-digoxygenin antibody coupled to a visible marker such as a fluorescent dye is used to visualize the probe. Other chemical labels such as biotin can be attached to nucleotides and used in essentially the same way.

prolonged period at 65°C. Similar hybridization reactions can occur between any two single-stranded nucleic acid chains (DNA/DNA, RNA/RNA, or RNA/DNA), provided that they have complementary nucleotide sequences. These specific hybridization reactions are widely used to detect and characterize specific nucleotide sequences in both RNA and DNA molecules.

Single-stranded DNA molecules used to detect complementary sequences are known as **probes**; these molecules, which carry radioactive or chemical markers to facilitate their detection, can range from fifteen to thousands of nucleotides long. Hybridization reactions using DNA probes are so sensitive and selective that they can detect complementary sequences present at a concentration as low as one molecule per cell. It is thus possible to determine how many copies of any DNA sequence are present in a particular DNA sample. The same technique can be used to search for related but nonidentical genes. To find a gene of interest in an organism whose genome has not yet been sequenced, for example, a portion of a known gene can be used as a probe (**Figure 8–36**).

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 194 of 240

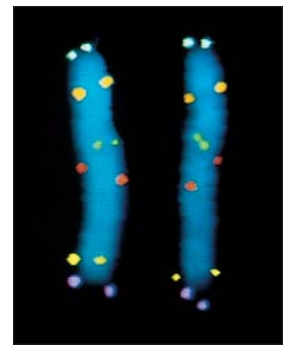


**Figure 8–35 *In situ* hybridization to locate specific genes on chromosomes.** Here, six different DNA probes have been used to mark the locations of their respective nucleotide sequences on human chromosome 5 at metaphase. The probes have been chemically labeled and detected with fluorescent antibodies. Both copies of chromosome 5 are shown, aligned side by side. Each probe produces two dots on each chromosome, since a metaphase chromosome has replicated its DNA and therefore contains two identical DNA helices. (Courtesy of David C. Ward.)

Alternatively, DNA probes can be used in hybridization reactions with RNA rather than DNA to find out whether a cell is expressing a given gene. In this case a DNA probe that contains part of the gene's sequence is hybridized with RNA purified from the cell in question to see whether the RNA includes nucleotide sequences matching the probe DNA and, if so, in what quantities. In somewhat more elaborate procedures, the DNA probe is treated with specific nucleases after the hybridization is complete, to determine the exact regions of the DNA probe that have paired with the RNA molecules. One can thereby determine the start and stop sites for RNA transcription, as well as the precise boundaries of the intron and exon sequences in a gene (**Figure 8–37**).

Today, the positions of intron/exon boundaries are usually determined by sequencing the *complementary DNA (cDNA)* sequences that represent the mRNAs expressed in a cell and comparing them with the nucleotide sequence of the genome. We describe later how cDNAs are prepared from mRNAs.

The hybridization of DNA probes to RNAs allows one to determine whether or not a particular gene is being transcribed; moreover, when the expression of a gene changes, one can determine whether the change is due to transcriptional or post-transcriptional controls (see Figure 7–92). These tests of gene expression were initially performed with one DNA probe at a time. *DNA microarrays* now allow the simultaneous monitoring of hundreds or thousands of genes at a time, as we discuss later. Hybridization methods are in such wide use in cell biology today that it is difficult to imagine how we could study gene structure and expression without them.

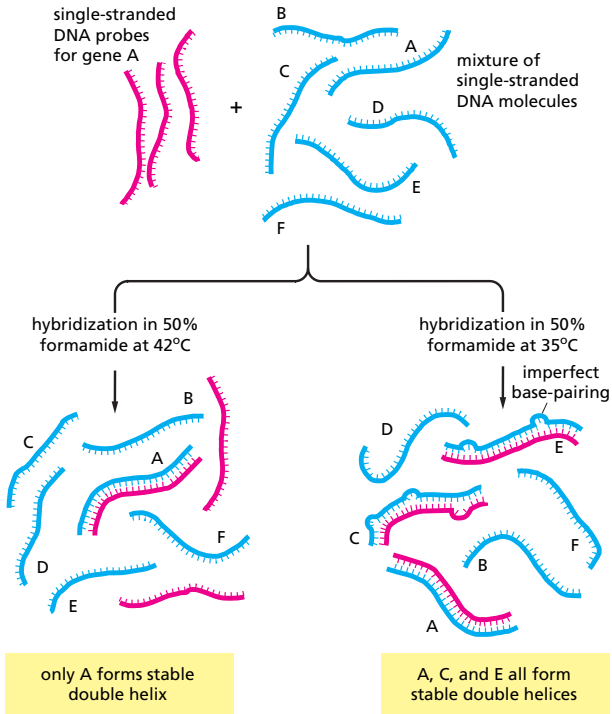



**Figure 8–36 Stringent versus nonstringent hybridization conditions.** To use a DNA probe to find an identical match, stringent hybridization conditions are used; the reaction temperature is kept just a few degrees below that at which a perfect DNA helix denatures in the solvent used (its *melting temperature*), so that all imperfect helices formed are unstable. When a DNA probe is being used to find DNAs with related, as well as identical, sequences, less stringent conditions are used; hybridization is performed at a lower temperature, which allows even imperfectly paired double helices to form. Only the lower-temperature hybridization conditions can be used to search for genes that are nonidentical but related to gene A (C and E in this example).

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 195 of 240

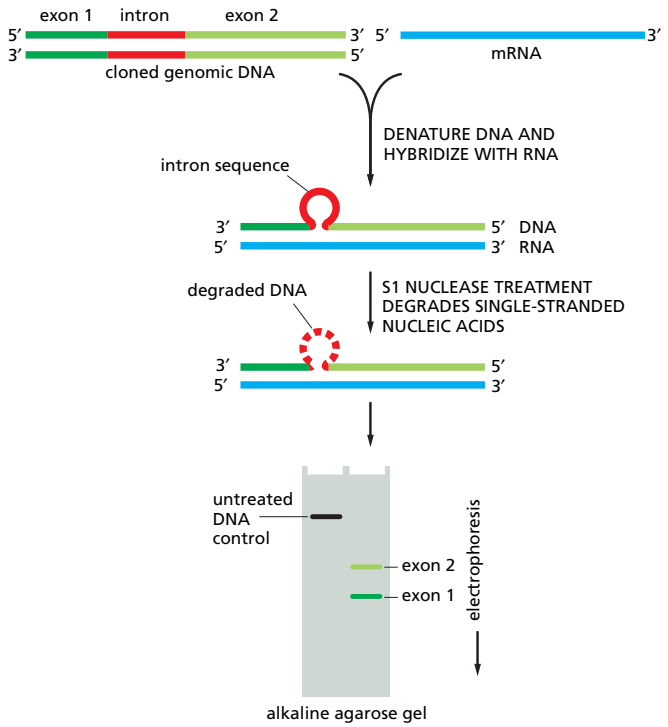



**Figure 8–37 The use of nucleic acid hybridization to determine the region of a cloned DNA fragment that is present in an mRNA molecule.** The method shown requires a nuclease that cuts the DNA chain only where it is not base-paired to a complementary RNA chain. The positions of the introns in eucaryotic genes are mapped by the method shown. For this type of analysis, the DNA is electrophoresed through a denaturing agarose gel, which causes it to migrate as single-stranded molecules. The location of each end of an RNA molecule can be determined using similar methods.

## Northern and Southern Blotting Facilitate Hybridization with Electrophoretically Separated Nucleic Acid Molecules

In a complex mixture of nucleic acids, DNA probes are often used to detect only those molecules with sequences that are complementary to all or part of the probe. Gel electrophoresis can be used to fractionate the many different RNA or DNA molecules in a crude mixture according to their size before the hybridization reaction is performed; if the probe binds to molecules of only one or a few sizes, one can be certain that the hybridization was indeed specific. Moreover, the size information obtained can be invaluable in itself. An example illustrates this point.

Suppose that one wishes to determine the nature of the defect in a mutant mouse that produces abnormally low amounts of albumin, a protein that liver cells normally secrete into the blood in large amounts. First, one collects identical samples of liver tissue from mutant and normal mice (the latter serving as controls) and disrupts the cells in a strong detergent to inactivate nucleases that might otherwise degrade the nucleic acids. Next, one separates the RNA and DNA from all of the other cell components: the proteins present are completely denatured and removed by repeated extractions with phenol—a potent organic solvent that is partly miscible with water; the nucleic acids, which remain in the aqueous phase, are then precipitated with alcohol to separate them from the small molecules of the cell. Then, one separates the DNA from the RNA by their different solubilities in alcohols and degrades any contaminating nucleic acid of the unwanted type by treatment with a highly specific enzyme—either an RNase or a DNase. The mRNAs are typically separated from bulk RNA by retention on a chromatography column that specifically binds the poly-A tails of mRNAs.

To analyze the albumin-encoding mRNAs, a technique called **Northern blotting** is used. First, the intact mRNA molecules purified from mutant and control liver cells are fractionated on the basis of their sizes into a series of bands by gel electrophoresis. Then, to make the RNA molecules accessible to DNA probes, a replica of the pattern of RNA bands on the gel is made by transferring ("blotting") the fractionated RNA molecules onto a sheet of nitrocellulose or nylon paper. The paper is then incubated in a solution containing a labeled DNA probe, the sequence of which corresponds to part of the template strand that

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 196 of 240

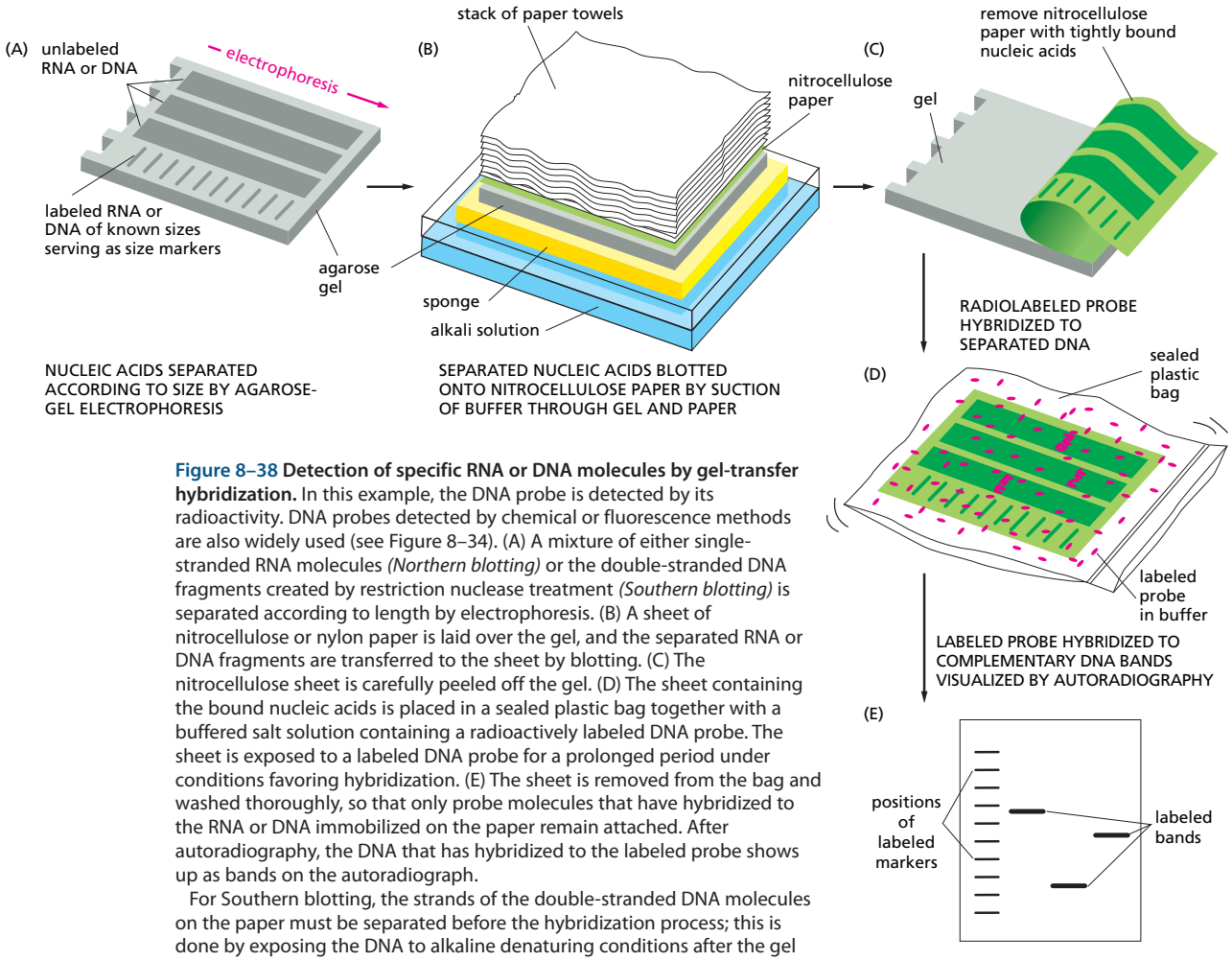



**Figure 8–38 Detection of specific RNA or DNA molecules by gel-transfer hybridization.** In this example, the DNA probe is detected by its radioactivity. DNA probes detected by chemical or fluorescence methods are also widely used (see Figure 8–34). (A) A mixture of either single-stranded RNA molecules *(Northern blotting)* or the double-stranded DNA fragments created by restriction nuclease treatment *(Southern blotting)* is separated according to length by electrophoresis. (B) A sheet of nitrocellulose or nylon paper is laid over the gel, and the separated RNA or DNA fragments are transferred to the sheet by blotting. (C) The nitrocellulose sheet is carefully peeled off the gel. (D) The sheet containing the bound nucleic acids is placed in a sealed plastic bag together with a buffered salt solution containing a radioactively labeled DNA probe. The sheet is exposed to a labeled DNA probe for a prolonged period under conditions favoring hybridization. (E) The sheet is removed from the bag and washed thoroughly, so that only probe molecules that have hybridized to the RNA or DNA immobilized on the paper remain attached. After autoradiography, the DNA that has hybridized to the labeled probe shows up as bands on the autoradiograph.

For Southern blotting, the strands of the double-stranded DNA molecules on the paper must be separated before the hybridization process; this is done by exposing the DNA to alkaline denaturing conditions after the gel has been run (not shown).

produces albumin mRNA. The RNA molecules that hybridize to the labeled DNA probe on the paper (because they are complementary to part of the normal albumin gene sequence) are then located by detecting the bound probe by autoradiography or by chemical means (**Figure 8–38**). The sizes of the hybridized RNA molecules can be determined by reference to RNA standards of known sizes that are electrophoresed side by side with the experimental sample. In this way, one might discover that liver cells from the mutant mice make albumin mRNA in normal amounts and of normal size; alternatively, you might find that they make it in normal size but in greatly reduced amounts. Another possibility is that the mutant albumin mRNA molecules are abnormally short; in this case the gel blot could be retested with a series of shorter DNA probes, each corresponding to small portions of the gene, to reveal which part of the normal RNA is missing.

The original gel-transfer hybridization method, called **Southern blotting**, analyzes DNA rather than RNA. (It was named after its inventor, and the Northern and Western blotting techniques were named with reference to it.) Here, isolated DNA is first cut into readily separable fragments with restriction nucleases. The double-stranded fragments are then separated on the basis of size by gel electrophoresis, and those complementary to a DNA probe are identified by blotting and hybridization, as just described for RNA (see Figure 8–38). To characterize the structure of the albumin gene in the mutant mice, an albumin-specific DNA probe would be used to construct a detailed *restriction map* of the genome in the region of the albumin gene (such a map consists of the pattern of DNA fragments produced by various restriction nucleases). From this map one

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 197 of 240

could determine if the albumin gene has been rearranged in the defective animals—for example, by the deletion or the insertion of a short DNA sequence; most single-base changes, however, could not be detected in this way.

## Genes Can Be Cloned Using DNA Libraries

Any DNA fragment can be cloned. In molecular biology, the term **DNA cloning** is used in two senses. In one sense, it literally refers to the act of making many identical copies of a DNA molecule—the amplification of a particular DNA sequence. However, the term also describes the isolation of a particular stretch of DNA (often a particular gene) from the rest of a cell's DNA, because this isolation is greatly facilitated by making many identical copies of the DNA of interest. As discussed earlier in this chapter, cloning, particularly when used in the context of developmental biology, can also refer to the generation of many genetically identical cells starting from a single cell or even to the generation of genetically identical organisms. In all cases, cloning refers to the act of making many genetically identical copies; in this section, we will use the term cloning (or DNA cloning or gene cloning) to refer to methods designed to generate many identical copies of a segment of nucleic acid.

DNA cloning in its most general sense can be accomplished in several ways. The simplest involves inserting a particular fragment of DNA into the purified DNA genome of a self-replicating genetic element—generally a virus or a plasmid. A DNA fragment containing a human gene, for example, can be joined in a test tube to the chromosome of a bacterial virus, and the new recombinant DNA molecule can then be introduced into a bacterial cell, where the inserted DNA fragment will be replicated along with the DNA of the virus. Starting with only one such recombinant DNA molecule that infects a single cell, the normal replication mechanisms of the virus can produce more than $10^{12}$ identical virus DNA molecules in less than a day, thereby amplifying the amount of the inserted human DNA fragment by the same factor. A virus or plasmid used in this way is known as a *cloning vector*, and the DNA propagated by insertion into it is said to have been *cloned*.

To isolate a specific gene, one often begins by constructing a *DNA library*—a comprehensive collection of cloned DNA fragments from a cell, tissue, or organism. This library includes (one hopes) at least one fragment that contains the gene of interest. Libraries can be constructed with either a virus or a plasmid vector and are generally housed in a population of bacterial cells. The principles underlying the methods used for cloning genes are the same for either type of cloning vector, although the details may differ. Today, most cloning is performed with plasmid vectors.

The **plasmid vectors** most widely used for gene cloning are small circular molecules of double-stranded DNA derived from larger plasmids that occur naturally in bacterial cells. They generally account for only a minor fraction of the total host bacterial cell DNA, but they can easily be separated owing to their small size from chromosomal DNA molecules, which are large and precipitate as a pellet upon centrifugation. For use as cloning vectors, the purified plasmid DNA circles are first cut with a restriction nuclease to create linear DNA molecules. The genomic DNA to be used in constructing the library is cut with the same restriction nuclease, and the resulting restriction fragments (including those containing the gene to be cloned) are then added to the cut plasmids and annealed via their cohesive ends to form recombinant DNA circles. These recombinant molecules containing foreign DNA inserts are then covalently sealed with the enzyme DNA ligase (**Figure 8–39**).

In the next step in preparing the library, the recombinant DNA circles are introduced into bacterial cells that have been made transiently permeable to DNA. These bacterial cells are now said to be *transfected* with the plasmids. As the cells grow and divide, doubling in number every 30 minutes, the recombinant plasmids also replicate to produce an enormous number of copies of DNA circles containing the foreign DNA (**Figure 8–40**). Many bacterial plasmids carry genes for antibiotic resistance (discussed in Chapter 24), a property that can be

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 198 of 240

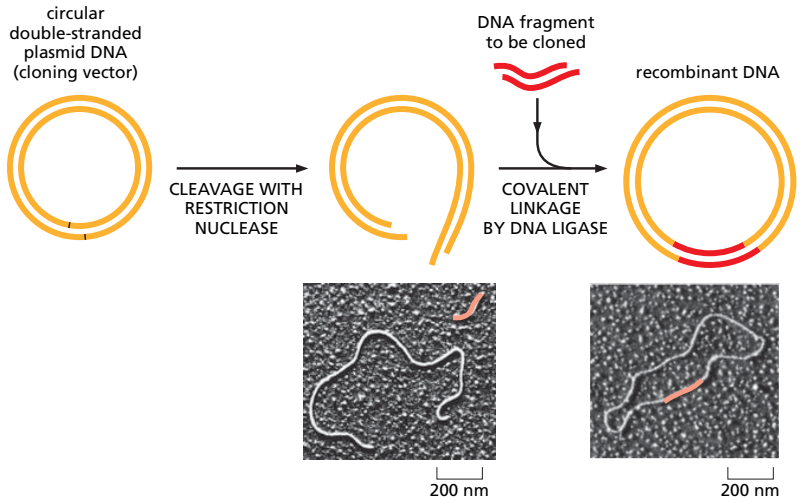



**Figure 8–39 The insertion of a DNA fragment into a bacterial plasmid with the enzyme DNA ligase.** The plasmid is cut open with a restriction nuclease (in this case one that produces cohesive ends) and is mixed with the DNA fragment to be cloned (which has been prepared with the same restriction nuclease). DNA ligase and ATP are added. The cohesive ends base-pair, and DNA ligase seals the nicks in the DNA backbone, producing a complete recombinant DNA molecule. (Micrographs courtesy of Huntington Potter and David Dressler.)

exploited to select those cells that have been successfully transfected; if the bacteria are grown in the presence of the antibiotic, only cells containing plasmids will survive. Each original bacterial cell that was initially transfected contains, in general, a different foreign DNA insert; this insert is inherited by all of the progeny cells of that bacterium, which together form a small colony in a culture dish.

For many years, plasmids were used to clone fragments of DNA of 1000 to 30,000 nucleotide pairs. Larger DNA fragments are more difficult to handle and were harder to clone. Then researchers began to use *yeast artificial chromosomes* (*YACs*), which could accommodate very large pieces of DNA (**Figure 8–41**). Today, new plasmid vectors based on the naturally occurring F plasmid of *E. coli* are used to clone DNA fragments of 300,000 to 1 million nucleotide pairs. Unlike smaller bacterial plasmids, the F plasmid—and its derivative, the **bacterial artificial chromosome** (**BAC**)—is present in only one or two copies per *E. coli* cell. The fact that BACs are kept in such low numbers in bacterial cells may contribute to their ability to maintain large cloned DNA sequences stably: with only a few BACs present, it is less likely that the cloned DNA fragments will become scrambled by recombination with sequences carried on other copies of the plasmid. Because of their stability, ability to accept large DNA inserts, and ease of handling, BACs are now the preferred vector for building DNA libraries of complex organisms—including those representing the human and mouse genomes.

## Two Types of DNA Libraries Serve Different Purposes

Cleaving the entire genome of a cell with a specific restriction nuclease and cloning each fragment as just described produces a very large number of DNA fragments—on the order of a million for a mammalian genome. The fragments are distributed among millions of different colonies of transfected bacterial cells.

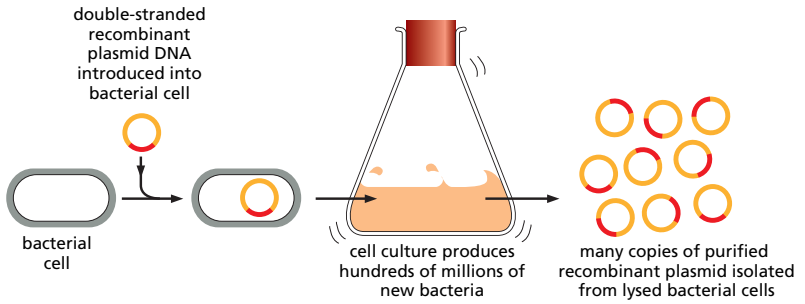



**Figure 8–40 The amplification of the DNA fragments inserted into a plasmid.** To produce large amounts of the DNA of interest, the recombinant plasmid DNA in Figure 8–39 is introduced into a bacterium by transfection, where it will replicate many millions of times as the bacterium multiplies.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 199 of 240

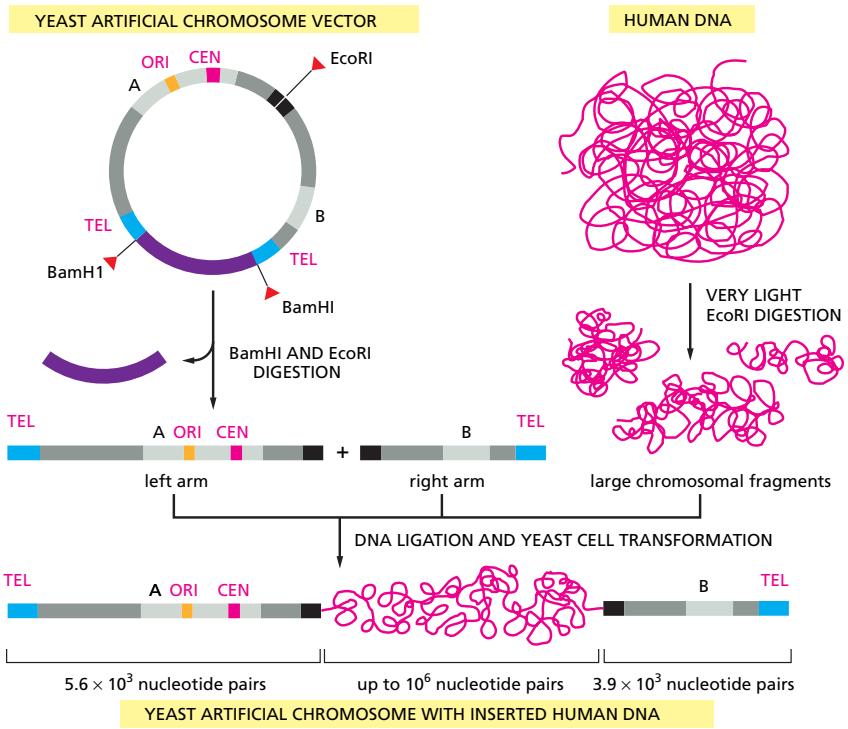



**Figure 8–41 The making of a yeast artificial chromosome (YAC).** A YAC vector allows the cloning of very large DNA molecules. TEL, CEN, and ORI are the telomere, centromere, and origin of replication sequences, respectively, for the yeast *Saccharomyces cerevisiae;* all of these are required to propagate the YAC. BamHI and EcoRI are sites where the corresponding restriction nucleases cut the DNA double helix. The sequences denoted A and B encode enzymes that serve as selectable markers to allow the easy isolation of yeast cells that have taken up the artificial chromosome. Because bacteria divide more rapidly than yeasts, most large-scale cloning projects now use *E. coli* as the means for amplifying DNA. (Adapted from D.T. Burke, G.F. Carle and M.V. Olson, *Science* 236:806–812, 1987. With permission from AAAS.)

When working with BACs rather than typical plasmids, larger fragments can be inserted, and so fewer transfected bacterial cells are required to cover the genome. In either case, each of the colonies is composed of a clone of cells derived from a single ancestor cell, and therefore harbors many copies of a particular stretch of the fragmented genome (**Figure 8–42**). Such a plasmid is said to contain a **genomic DNA clone**, and the entire collection of plasmids is called a **genomic DNA library**. But because the genomic DNA is cut into fragments at random, only some fragments contain genes. Many of the genomic DNA clones obtained from the DNA of a higher eucaryotic cell contain only noncoding DNA, which, as we discussed in Chapter 4, makes up most of the DNA in such genomes.

An alternative strategy is to begin the cloning process by selecting only those DNA sequences that are transcribed into mRNA and thus are presumed to correspond to protein-encoding genes. This is done by extracting the mRNA from cells and then making a DNA copy of each mRNA molecule present—a so-called *complementary DNA*, or *cDNA*. The copying reaction is catalyzed by the reverse transcriptase enzyme of retroviruses, which synthesizes a complementary DNA chain on an RNA template. The single-stranded cDNA molecules synthesized by the reverse transcriptase are converted into double-stranded cDNA molecules by DNA polymerase, and these molecules are inserted into a plasmid or virus vector and cloned (**Figure 8–43**). Each clone obtained in this way is called a **cDNA clone**, and the entire collection of clones derived from one mRNA preparation constitutes a **cDNA library**.

**Figure 8–44** illustrates some important differences between genomic DNA clones and cDNA clones. Genomic clones represent a random sample of all of the DNA sequences in an organism and, with very rare exceptions, are the same regardless of the cell type used to prepare them. By contrast, cDNA clones contain only those regions of the genome that have been transcribed into mRNA. Because the cells of different tissues produce distinct sets of mRNA molecules, a distinct cDNA library is obtained for each type of cell used to prepare the library.

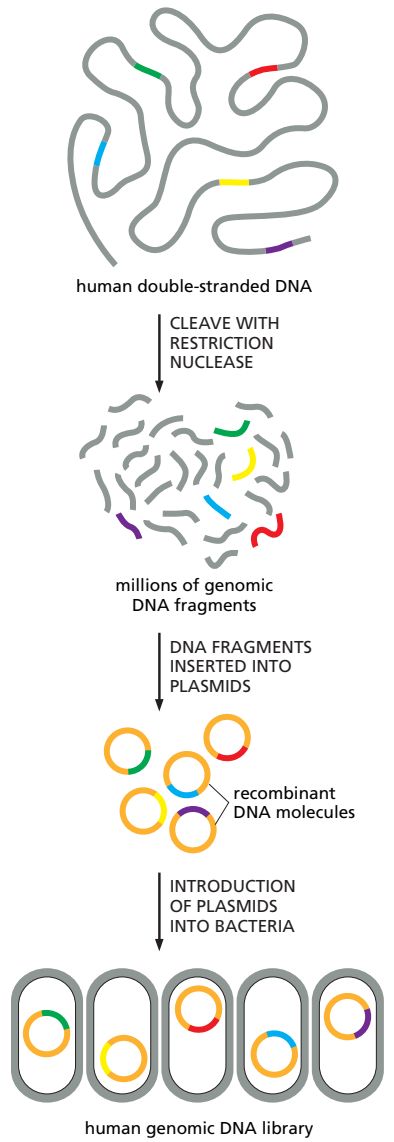



**Figure 8–42 Construction of a human genomic DNA library.** A genomic library is usually stored as a set of bacteria, each bacterium carrying a different fragment of human DNA. For simplicity, cloning of just a few representative fragments *(colored)* is shown. In reality, all of the *gray* DNA fragments would also be cloned.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 200 of 240

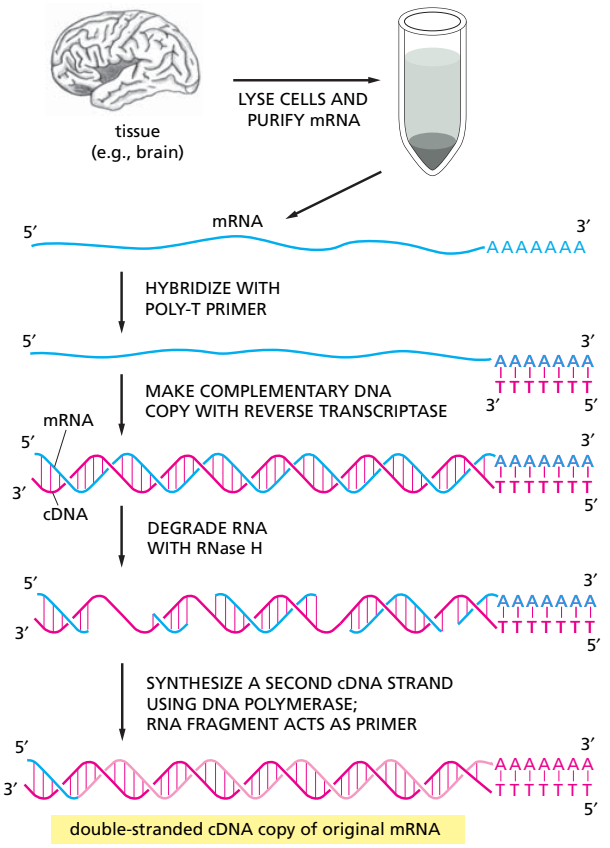



**Figure 8–43 The synthesis of cDNA.** Total mRNA is extracted from a particular tissue, and the enzyme reverse transcriptase produces DNA copies (cDNA) of the mRNA molecules (see p. 320). For simplicity, the copying of just one of these mRNAs into cDNA is illustrated. A short oligonucleotide complementary to the poly-A tail at the 3′ end of the mRNA (discussed in Chapter 6) is first hybridized to the RNA to act as a primer for the reverse transcriptase, which then copies the RNA into a complementary DNA chain, thereby forming a DNA/RNA hybrid helix. Treating the DNA/RNA hybrid with RNase H (see Figure 5–12) creates nicks and gaps in the RNA strand. The enzyme DNA polymerase then copies the remaining single-stranded cDNA into double-stranded cDNA. The fragment of the original mRNA is the primer for this synthesis reaction, as shown. Because the DNA polymerase used to synthesize the second DNA strand can synthesize through the bound RNA molecules, the RNA fragment that is base-paired to the 3′ end of the first DNA strand usually acts as the primer for the final product of the second strand synthesis. This RNA is eventually degraded during subsequent cloning steps. As a result, the nucleotide sequences at the extreme 5′ ends of the original mRNA molecules are often absent from cDNA libraries.

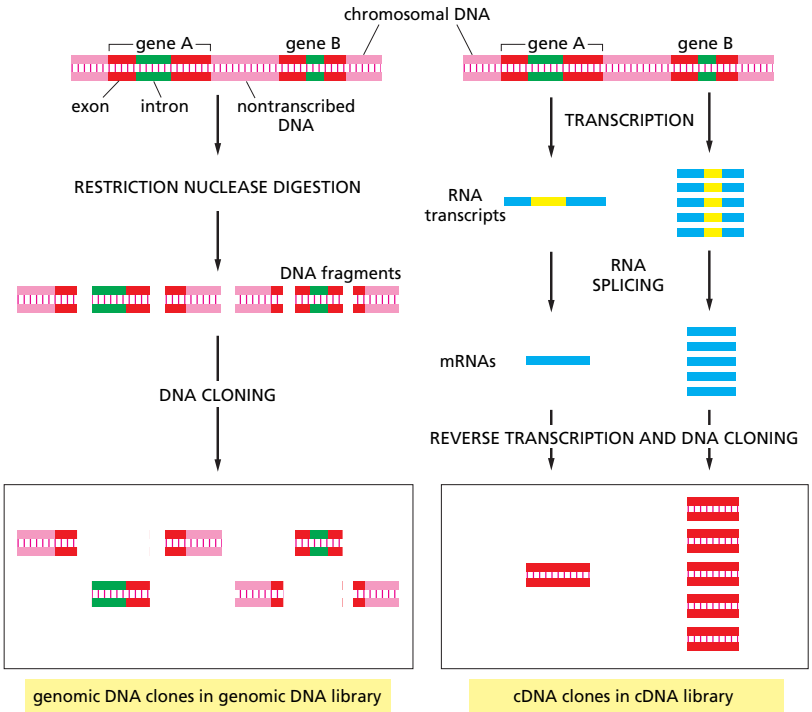



**Figure 8–44 The differences between cDNA clones and genomic DNA clones derived from the same region of DNA.** In this example, gene A is infrequently transcribed, whereas gene B is frequently transcribed, and both genes contain introns *(green)*. In the genomic DNA library, both the introns and the nontranscribed DNA *(pink)* are included in the clones, and most clones contain, at most, only part of the coding sequence of a gene *(red)*. In the cDNA clones, the intron sequences *(yellow)* have been removed by RNA splicing during the formation of the mRNA *(blue)*, and a continuous coding sequence is therefore present in each clone. Because gene B is transcribed more frequently than gene A in the cells from which the cDNA library was made, it is represented much more frequently than A in the cDNA library. In contrast, A and B are in principle represented equally in the genomic DNA library.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 201 of 240

## cDNA Clones Contain Uninterrupted Coding Sequences

There are several advantages in using a cDNA library for gene cloning. First, specialized cells produce large quantities of some proteins. In this case, the mRNA encoding the protein is likely to be produced in such large quantities that a cDNA library prepared from the cells is highly enriched for the cDNA molecules encoding the protein, greatly reducing the problem of identifying the desired clone in the library (see Figure 8–44). Hemoglobin, for example, is made in large amounts by developing erythrocytes (red blood cells); for this reason the globin genes were among the first to be cloned.

By far the most important advantage of cDNA clones is that they contain the uninterrupted coding sequence of a gene. As we have seen, eucaryotic genes usually consist of short coding sequences of DNA (exons) separated by much longer noncoding sequences (introns); the production of mRNA entails the removal of the noncoding sequences from the initial RNA transcript and the splicing together of the coding sequences. Neither bacterial nor yeast cells will make these modifications to the RNA produced from a gene of a higher eucaryotic cell. Thus, when the aim of the cloning is either to deduce the amino acid sequence of the protein from the DNA sequence or to produce the protein in bulk by expressing the cloned gene in a bacterial or yeast cell, it is much preferable to start with cDNA. cDNA libraries have an additional use: as described in Chapter 7, many mRNAs from humans and other complex organisms are alternatively spliced, and a cDNA library often represents many, if not all, of the alternatively spliced mRNAs produced from a given cell line or tissue.

Genomic and cDNA libraries are inexhaustible resources, which are widely shared among investigators. Today, many such libraries are also available from commercial sources.

## Genes Can Be Selectively Amplified by PCR

Now that so many genome sequences are available, genes can be cloned directly without the need to first construct DNA libraries. A technique called the **polymerase chain reaction** (**PCR**) makes this rapid cloning possible. Starting with an entire genome, PCR allows the DNA from a selected region to be amplified several billionfold, effectively "purifying" this DNA away from the remainder of the genome.

To begin, a pair of DNA oligonucleotides, chosen to flank the desired nucleotide sequence of the gene, are synthesized by chemical methods. These oligonucleotides are then used to prime DNA synthesis on single strands generated by heating the DNA from the entire genome. The newly synthesized DNA is produced in a reaction catalyzed *in vitro* by a purified DNA polymerase, and the primers remain at the 5′ ends of the final DNA fragments that are made (**Figure 8–45**A).

Nothing special is produced in the first cycle of DNA synthesis; the power of the PCR method is revealed only after repeated rounds of DNA synthesis. Every cycle doubles the amount of DNA synthesized in the previous cycle. Because each cycle requires a brief heat treatment to separate the two strands of the template DNA double helix, the technique requires the use of a special DNA polymerase, isolated from a thermophilic bacterium, that is stable at much higher temperatures than normal so that it is not denatured by the repeated heat treatments. With each round of DNA synthesis, the newly generated fragments serve as templates in their turn, and within a few cycles the predominant product is a single species of DNA fragment whose length corresponds to the distance between the two original primers (see Figure 8–45B).

In practice, effective DNA amplification requires 20–30 reaction cycles, with the products of each cycle serving as the DNA templates for the next—hence the term polymerase "chain reaction." A single cycle requires only about 5 minutes, and the entire procedure can be easily automated. PCR thereby makes possible the "cell-free molecular cloning" of a DNA fragment in a few hours, compared with the several days required for standard cloning procedures. This technique

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 202 of 240

is now used routinely to clone DNA from genes of interest directly—starting either from genomic DNA or from mRNA isolated from cells (**Figure 8–46**).

The PCR method is extremely sensitive; it can detect a single DNA molecule in a sample. Trace amounts of RNA can be analyzed in the same way by first transcribing them into DNA with reverse transcriptase. The PCR cloning technique has largely replaced Southern blotting for the diagnosis of genetic diseases and for the detection of low levels of viral infection. It also has great promise in forensic medicine as a means of analyzing minute traces of blood or other tissues—

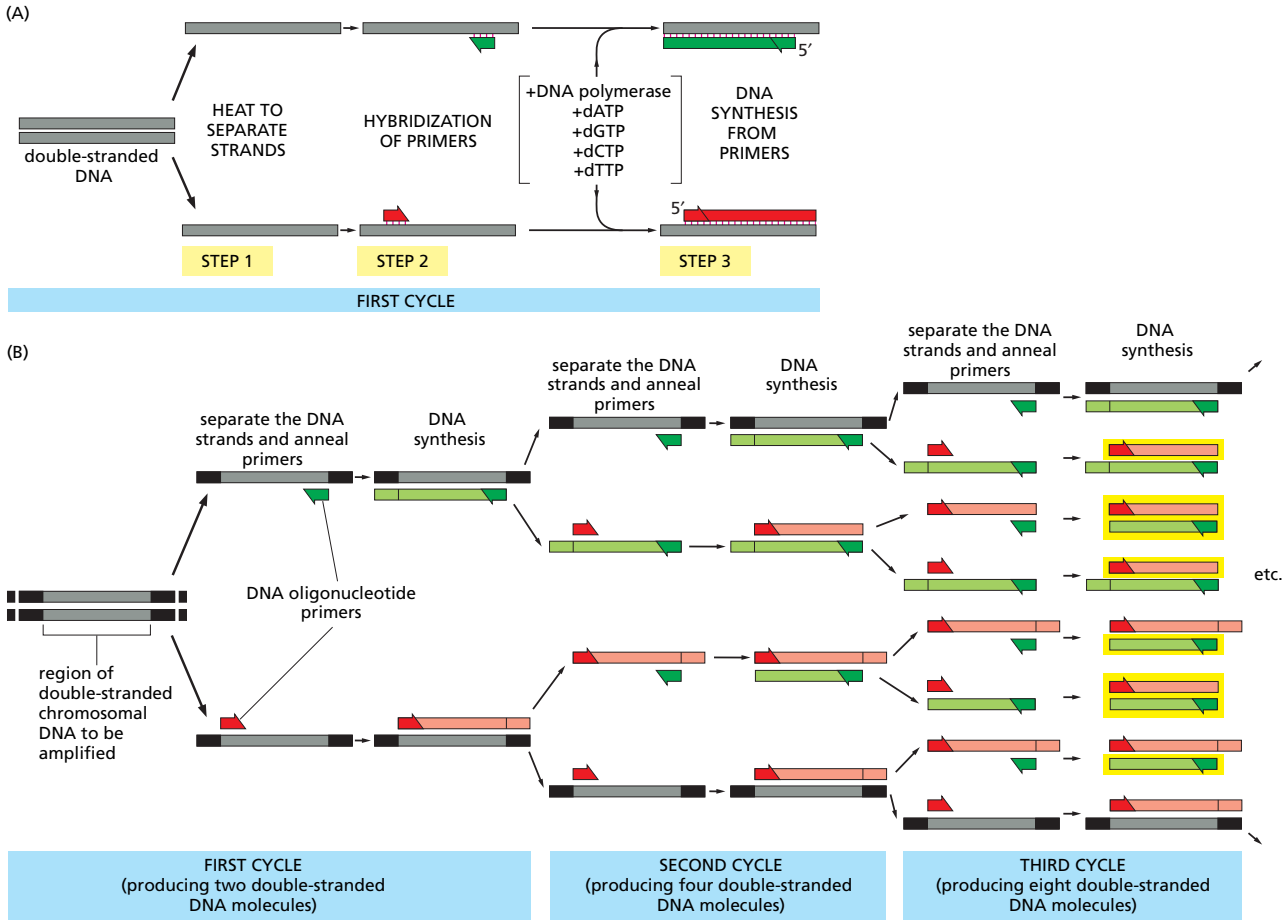



**Figure 8–45 Amplification of DNA by the PCR technique.** <TACG> Knowledge of the DNA sequence to be amplified is used to design two synthetic, primer DNA oligonucleotides. One primer is complementary to the sequence on one strand of the DNA double helix, and one is complementary to the sequence on the other strand, but at the opposite end of the region to be amplified. These oligonucleotides serve as primers for *in vitro* DNA synthesis, which is performed by a DNA polymerase, and they determine the segment of the DNA to be amplified. (A) PCR starts with a double-stranded DNA, and each cycle of the reaction begins with a brief heat treatment to separate the two strands (step 1). After strand separation, cooling of the DNA in the presence of a large excess of the two primer DNA oligonucleotides allows these primers to hybridize to complementary sequences in the two DNA strands (step 2). This mixture is then incubated with DNA polymerase and the four deoxyribonucleoside triphosphates to synthesize DNA, starting from the two primers (step 3). The entire cycle is then begun again by a heat treatment to separate the newly synthesized DNA strands. (B) As the procedure is performed over and over again, the newly synthesized fragments serve as templates in their turn, and within a few cycles the predominant DNA is identical to the sequence bracketed by and including the two primers in the original template. Of the DNA put into the original reaction, only the sequence bracketed by the two primers is amplified because there are no primers attached anywhere else. In the example illustrated in (B), three cycles of reaction produce 16 DNA chains, 8 of which *(boxed in yellow)* are the same length as and correspond exactly to one or the other strand of the original bracketed sequence shown at the far left; the other strands contain extra DNA downstream of the original sequence, which is replicated in the first few cycles. After four more cycles, 240 of the 256 DNA chains correspond exactly to the original bracketed sequence, and after several more cycles, essentially all of the DNA strands have this unique length.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 203 of 240

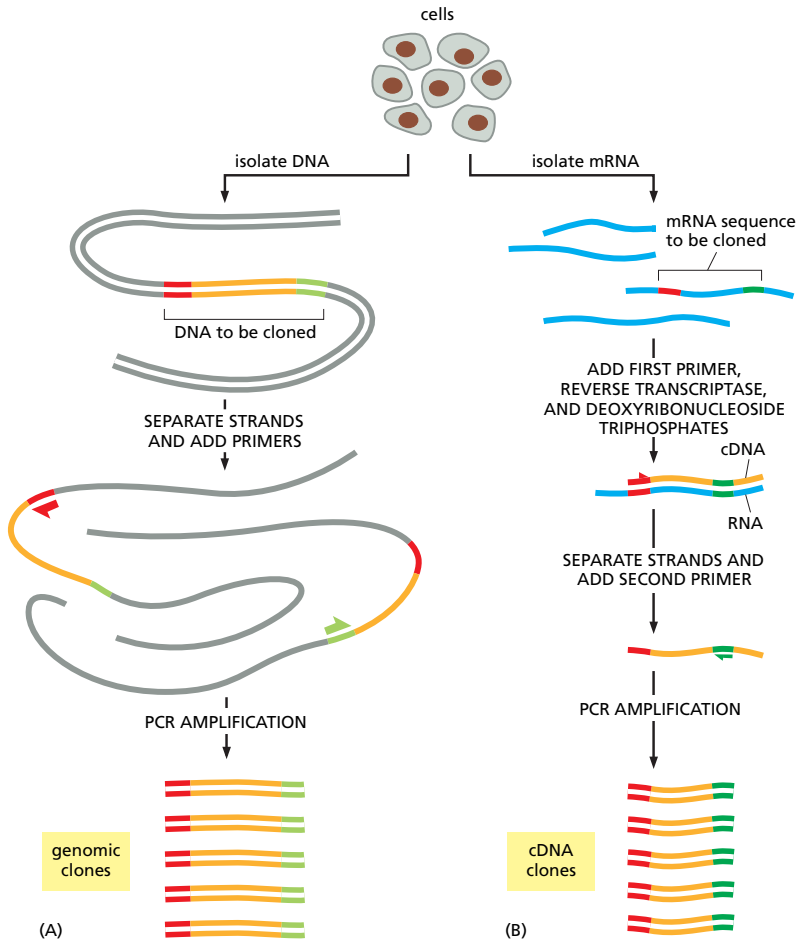



**Figure 8–46 Use of PCR to obtain a genomic or cDNA clone.** (A) To obtain a genomic clone using PCR, chromosomal DNA is first purified from cells. PCR primers that flank the stretch of DNA to be cloned are added, and many cycles of the reaction are completed (see Figure 8–45). Since only the DNA between (and including) the primers is amplified, PCR provides a way to obtain a short stretch of chromosomal DNA selectively in a virtually pure form. (B) To use PCR to obtain a cDNA clone of a gene, mRNA is first purified from cells. The first primer is then added to the population of mRNAs, and reverse transcriptase is used to make a complementary DNA strand. The second primer is then added, and the single-stranded cDNA molecule is amplified through many cycles of PCR, as shown in Figure 8–45. For both types of cloning, the nucleotide sequence of at least part of the region to be cloned must be known beforehand.

even as little as a single cell—and identifying the person from whom the sample came by his or her genetic "fingerprint" (**Figure 8–47**).

## Cells Can Be Used As Factories to Produce Specific Proteins

The vast majority of the thousands of different proteins in a cell, including many with crucially important functions, are present in very small amounts. In the past, for most of them, it has been extremely difficult, if not impossible, to obtain more than a few micrograms of pure material. One of the most important contributions of DNA cloning and genetic engineering to cell biology is that they have made it possible to produce any of the cell's proteins in nearly unlimited amounts.

Large amounts of a desired protein are produced in living cells by using **expression vectors** (**Figure 8–48**). These are generally plasmids that have been designed to produce a large amount of a stable mRNA that can be efficiently translated into protein in the transfected bacterial, yeast, insect, or mammalian cell. To prevent the high level of the foreign protein from interfering with the transfected cell's growth, the expression vector is often designed to delay the synthesis of the foreign mRNA and protein until shortly before the cells are harvested and lysed (**Figure 8–49**).

Because the desired protein made from an expression vector is produced inside a cell, it must be purified away from the host-cell proteins by chromatography after cell lysis; but because it is such a plentiful species in the cell lysate (often 1–10% of the total cell protein), the purification is usually easy to accomplish in only a few steps. As we saw above, many expression vectors have been

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 204 of 240

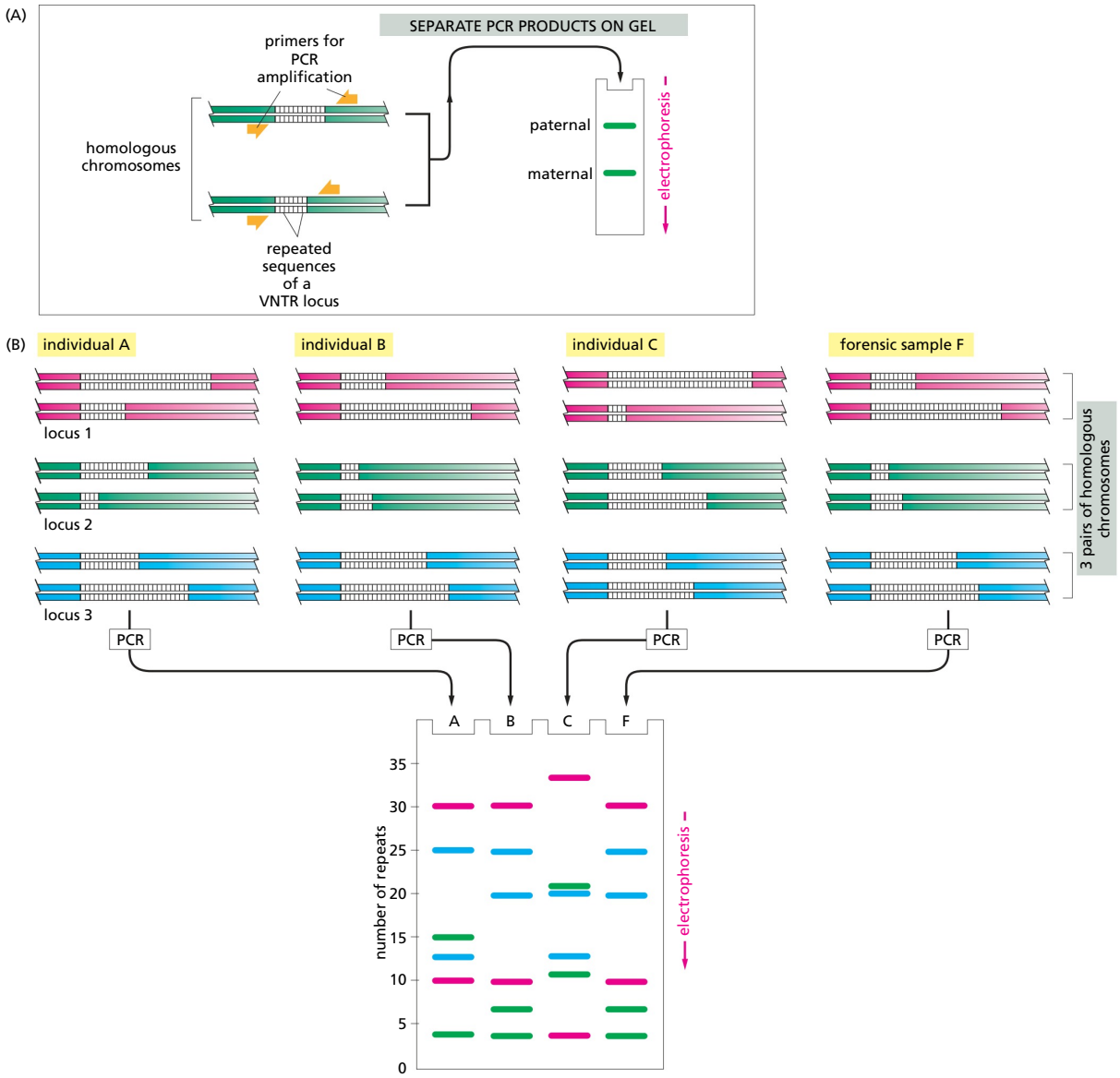



**Figure 8–47 How PCR is used in forensic science.** (A) The DNA sequences that create the variability used in this analysis contain runs of short, repeated sequences, such as CACACA . . ., which are found in various positions (loci) in the human genome. The number of repeats in each run can be highly variable in the population, ranging from 4 to 40 in different individuals. A run of repeated nucleotides of this type is commonly referred to as a *hypervariable microsatellite* sequence—also known as a VNTR *(variable number of tandem repeat)* sequence. Because of the variability in these sequences at each locus, individuals usually inherit a different variant from their mother and from their father; two unrelated individuals therefore do not usually contain the same pair of sequences. A PCR analysis using primers that bracket the locus produces a pair of bands of amplified DNA from each individual, one band representing the maternal variant and the other representing the paternal variant. The length of the amplified DNA, and thus the position of the band it produces after electrophoresis, depends on the exact number of repeats at the locus. (B) In the schematic example shown here, the same three VNTR loci are analyzed (requiring three different pairs of specially selected oligonucleotide primers) from three suspects (individuals A, B, and C), producing six DNA bands for each person after polyacrylamide-gel electrophoresis. Although some individuals have several bands in common, the overall pattern is quite distinctive for each. The band pattern can therefore serve as a "fingerprint" to identify an individual nearly uniquely. The fourth lane (F) contains the products of the same reactions carried out on a forensic sample. The starting material for such a PCR can be a single hair or a tiny sample of blood that was left at the crime scene. When examining the variability at 5–10 different VNTR loci, the odds that two random individuals would share the same genetic pattern by chance can be approximately 1 in 10 billion. In the case shown here, individuals A and C can be eliminated from further enquiries, whereas individual B remains a clear suspect for committing the crime. A similar approach is now routinely used for paternity testing.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 205 of 240

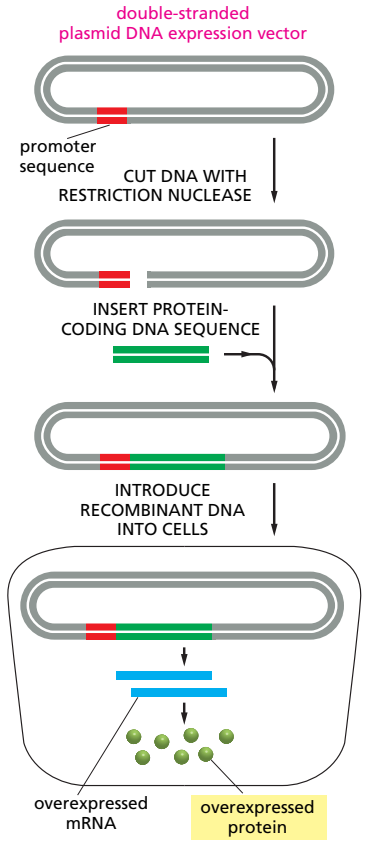



**Figure 8–48 Production of large amounts of a protein from a protein-coding DNA sequence cloned into an expression vector and introduced into cells.** A plasmid vector has been engineered to contain a highly active promoter, which causes unusually large amounts of mRNA to be produced from an adjacent protein-coding gene inserted into the plasmid vector. Depending on the characteristics of the cloning vector, the plasmid is introduced into bacterial, yeast, insect, or mammalian cells, where the inserted gene is efficiently transcribed and translated into protein.

designed to add a molecular tag—a cluster of histidine residues or a small marker protein—to the expressed protein to allow easy purification by affinity chromatography (see Figure 8–16). A variety of expression vectors are available, each engineered to function in the type of cell in which the protein is to be made. In this way, cells can be induced to make vast quantities of medically useful proteins—such as human insulin and growth hormone, interferon, and viral antigens for vaccines. More generally, these methods make it possible to produce every protein—even those that may be present in only a few copies per cell—in large enough amounts to be used in the kinds of detailed structural and functional studies that we discussed earlier.

DNA technology also can produce large amounts of any RNA molecule whose gene has been isolated. Studies of RNA splicing, protein synthesis, and RNA-based enzymes, for example, are greatly facilitated by the availability of pure RNA molecules. Most RNAs are present in only tiny quantities in cells, and they are very difficult to purify away from other cell components—especially from the many thousands of other RNAs present in the cell. But any RNA of interest can be synthesized efficiently *in vitro* by transcription of its DNA sequence (produced by one of the methods just described) with a highly efficient viral RNA polymerase. The single species of RNA produced is then easily purified away from the DNA template and the RNA polymerase.

## Proteins and Nucleic Acids Can Be Synthesized Directly by Chemical Reactions

Chemical reactions have been devised to synthesize directly specific sequences of amino acids or nucleic acids. These methodologies provide direct sources of biological molecules and do not rely on any cells or enzymes. Chemical synthesis is the method of choice for obtaining nucleic acids in the range of 100 nucleotides or fewer, which are particularly useful in the PCR-based approaches discussed above. Chemical synthesis is also routinely used to produce specific peptides that, when chemically coupled to other proteins, are used to generate antibodies against the peptide.

## DNA Can Be Rapidly Sequenced

Methods that allow the nucleotide sequence of any DNA fragment to be determined simply and quickly have made it possible to determine the DNA sequences of tens of thousands of genes, and many complete genomes (see Table 1–1, p. 18). The volume of DNA sequence information is now so large (many tens of billions of nucleotides) that powerful computers must be used to store and analyze it.

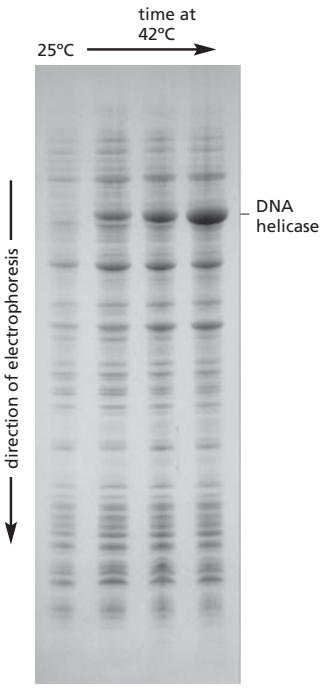



**Figure 8–49 Production of large amounts of a protein by using a plasmid expression vector.** In this example, bacterial cells have been transfected with the coding sequence for an enzyme, DNA helicase; transcription from this coding sequence is under the control of a viral promoter that becomes active only at temperatures of 37°C or higher. The total cell protein has been analyzed by SDS polyacrylamide-gel electrophoresis, either from bacteria grown at 25°C (no helicase protein made) or after a shift of the same bacteria to 42°C for up to 2 hours (helicase protein has become the most abundant protein species in the lysate). (Courtesy of Jack Barry.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 206 of 240

Large-volume DNA sequencing was made possible through the development in the mid-1970s of the **dideoxy method** for sequencing DNA, which is based on *in vitro* DNA synthesis performed in the presence of chain-terminating dideoxyribonucleoside triphosphates (**Figure 8–50**).

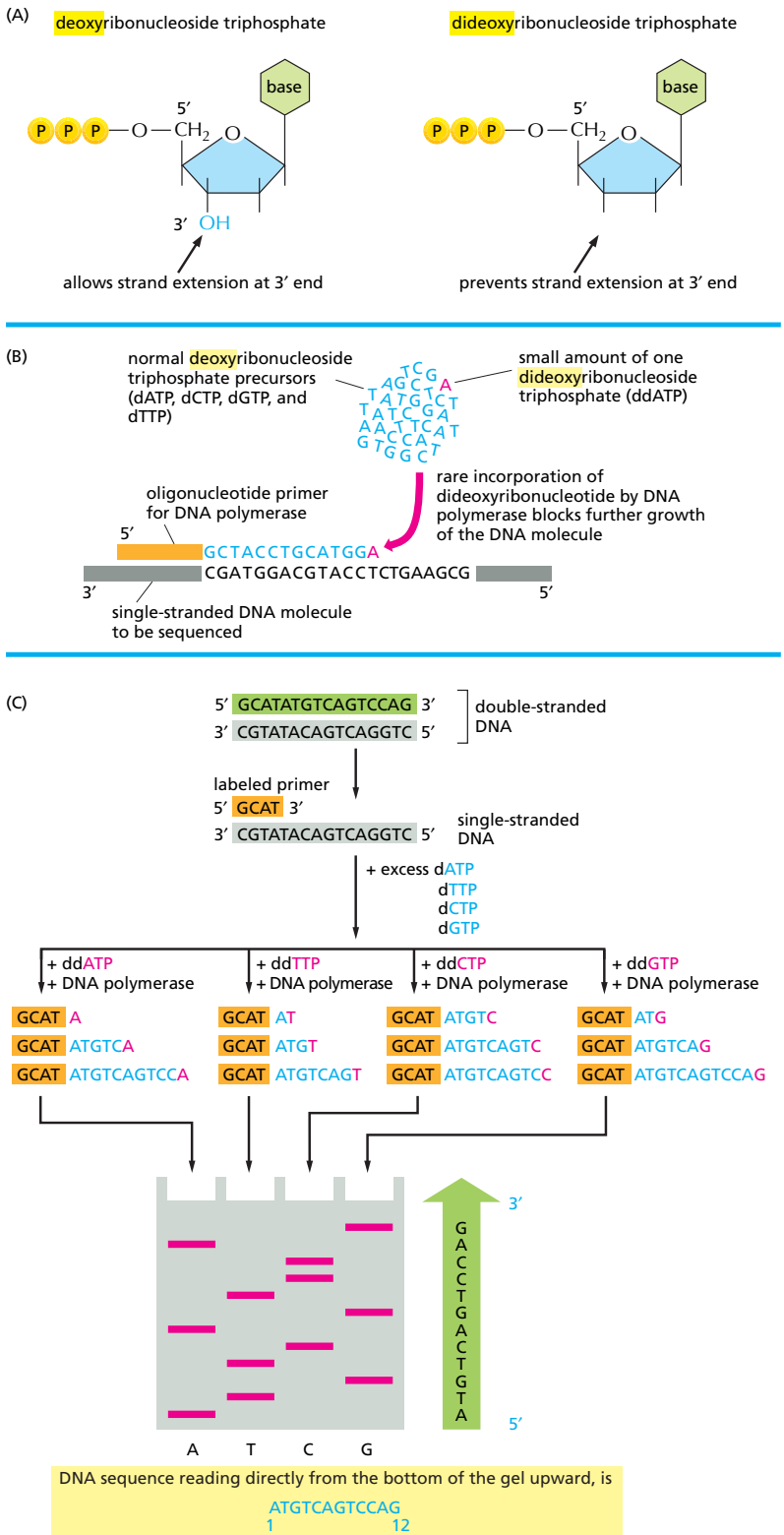



**Figure 8–50 The enzymatic—or dideoxy—method of sequencing DNA.** (A) This method relies on the use of dideoxyribonucleoside triphosphates, derivatives of the normal deoxyribonucleoside triphosphates that lack the 3′ hydroxyl group. (B) Purified DNA is synthesized *in vitro* in a mixture that contains single-stranded molecules of the DNA to be sequenced *(gray),* the enzyme DNA polymerase, a short primer DNA *(orange)* to enable the polymerase to start DNA synthesis, and the four deoxyribonucleoside triphosphates (dATP, dCTP, dGTP, dTTP: *blue* A, C, G, and T). If a dideoxyribonucleotide analog *(red)* of one of these nucleotides is also present in the nucleotide mixture, it can become incorporated into a growing DNA chain. Because this chain now lacks a 3′ OH group, the addition of the next nucleotide is blocked, and the DNA chain terminates at that point. In the example illustrated, a small amount of dideoxyATP (ddATP, symbolized here as a *red* A) has been included in the nucleotide mixture. It competes with an excess of the normal deoxyATP (dATP, *blue* A), so that ddATP is occasionally incorporated, at random, into a growing DNA strand. This reaction mixture will eventually produce a set of DNAs of different lengths complementary to the template DNA that is being sequenced and terminating at each of the different As. The exact lengths of the DNA synthesis products can then be used to determine the position of each A in the growing chain. (C) To determine the complete sequence of a DNA fragment, the double-stranded DNA is first separated into its single strands and one of the strands is used as the template for sequencing. Four different chain-terminating dideoxyribonucleoside triphosphates (ddATP, ddCTP, ddGTP, ddTTP, again shown in *red)* are used in four separate DNA synthesis reactions on copies of the same single-stranded DNA template *(gray).* Each reaction produces a set of DNA copies that terminate at different points in the sequence. The products of these four reactions are separated by electrophoresis in four parallel lanes of a polyacrylamide gel (labeled here A, T, C, and G). The newly synthesized fragments are detected by a label (either radioactive or fluorescent) that has been incorporated either into the primer or into one of the deoxyribonucleoside triphosphates used to extend the DNA chain. In each lane, the bands represent fragments that have terminated at a given nucleotide (e.g., A in the leftmost lane) but at different positions in the DNA. By reading off the bands in order, starting at the bottom of the gel and working across all lanes, the DNA sequence of the newly synthesized strand can be determined. The sequence is given in the *green arrow* to the right of the gel. This sequence is complementary to the template strand *(gray)* from the original double-stranded DNA molecule, and identical to a portion of the *green* 5′-to-3′ strand.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 207 of 240

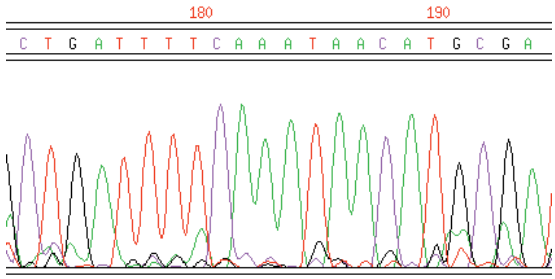



**Figure 8–51 Automated DNA sequencing.** Shown at the bottom is a tiny part of the raw data from an automated DNA-sequencing run as it appears on the computer screen. Each prominant colored peak represents a nucleotide in the DNA sequence—a clear stretch of nucleotide sequence can be read here between positions 173 and 194 from the start of the sequence. The small peaks along the baseline represent background "noise" and, as long as they are much lower than the "signal" peaks, they are ignored. This particular example is taken from the international project that determined the complete nucleotide sequence of the genome of the plant *Arabidopsis*. (Courtesy of George Murphy.)

Although the same basic method is still used today, many improvements have been made. DNA sequencing is now completely automated: robotic devices mix the reagents and then load, run, and read the order of the nucleotide bases from the gel. Chain-terminating nucleotides that are each labeled with a different colored fluorescent dye facilitate these tasks; in this case, all four synthesis reactions can be performed in the same tube, and the products can be separated in a single lane of a gel. A detector positioned near the bottom of the gel reads and records the color of the fluorescent label on each band as it passes through a laser beam (**Figure 8–51**). A computer then reads and stores this nucleotide sequence. Some modern systems dispense with the traditional gel entirely, separating nucleic acids by capillary electrophoresis, a method that facilitates rapid automation.

## Nucleotide Sequences Are Used to Predict the Amino Acid Sequences of Proteins

Now that DNA sequencing is so rapid and reliable, it has become the preferred method for determining, indirectly, the amino acid sequences of most proteins. Given a nucleotide sequence that encodes a protein, the procedure is quite straightforward. Although in principle there are six different reading frames in which a DNA sequence can be translated into protein (three on each strand), the correct one is generally recognizable as the only one lacking frequent stop codons (**Figure 8–52**). As we saw when we discussed the genetic code in Chap-

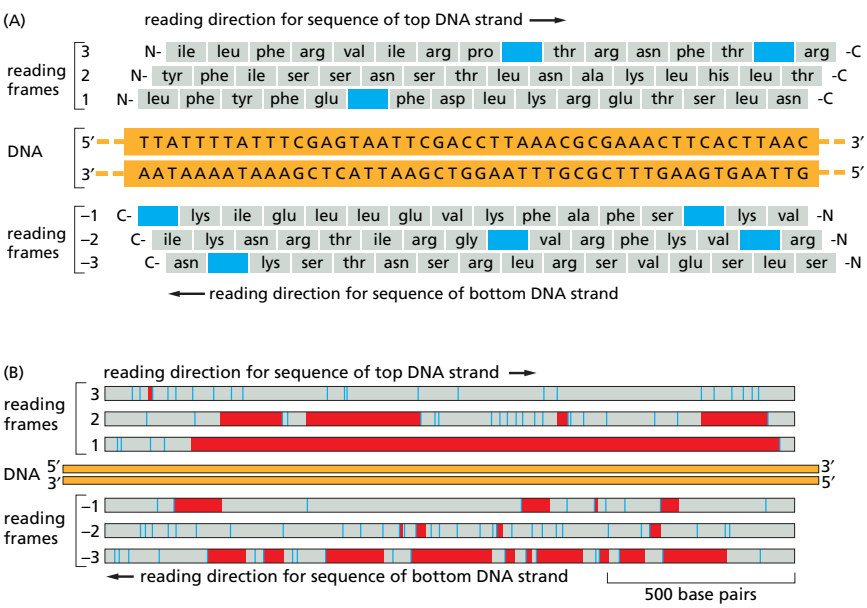



**Figure 8–52 Finding the regions in a DNA sequence that encode a protein.** (A) Any region of the DNA sequence can, in principle, code for six different amino acid sequences, because any one of three different reading frames can be used to interpret the nucleotide sequence on each strand. Note that a nucleotide sequence is always read in the 5′-to-3′ direction and encodes a polypeptide from the N-terminus to the C-terminus. For a random nucleotide sequence read in a particular frame, a stop signal for protein synthesis is encountered, on average, about once every 20 amino acids. In this sample sequence of 48 base pairs, each such signal *(stop codon)* is colored *blue*, and only reading frame 2 lacks a stop signal. (B) Search of a 1700 base-pair DNA sequence for a possible protein-encoding sequence. The information is displayed as in (A), with each stop signal for protein synthesis denoted by a *blue line*. In addition, all of the regions between possible start and stop signals for protein synthesis (see p. 381) are displayed as *red bars*. Only reading frame 1 actually encodes a protein, which is 475 amino acid residues long.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 208 of 240

ter 6, a random sequence of nucleotides, read in frame, will encode a stop signal for protein synthesis about once every 20 amino acids. Nucleotide sequences that encode a stretch of amino acids much longer than this are candidates for presumptive exons, and they can be translated (by computer) into amino acid sequences and checked against databases for similarities to known proteins from other organisms. If necessary, a limited amount of amino acid sequence can then be determined from the purified protein to confirm the sequence predicted from the DNA.

The problem comes, however, in determining which nucleotide sequences—within a whole genome—represent genes that encode proteins. Identifying genes is easiest when the DNA sequence is from a bacterial or archaeal chromosome, which lacks introns, or from a cDNA clone. The location of genes in these nucleotide sequences can be predicted by examining the DNA for certain distinctive features (discussed in Chapter 6). Briefly, these genes that encode proteins are identified by searching the nucleotide sequence for *open reading frames (ORFs)* that begin with an initiation codon, usually ATG, and end with a termination codon, TAA, TAG, or TGA. To minimize errors, computers used to search for ORFs are often directed to count as genes only those sequences that are longer than, say, 100 codons in length.

For more complex genomes, such as those of animals and plants, the presence of large introns embedded within the coding portion of genes complicates the process. In many multicellular organisms, including humans, the average exon is only 150 nucleotides long. Thus one must also search for other features that signal the presence of a gene, for example, sequences that signal an intron/exon boundary or distinctive upstream regulatory regions. Recent efforts to solve the exon prediction problem have turned to artificial intelligence algorithms, in which the computer learns, based on known examples, what sets of features are most indicative of an exon boundary.

A second major approach to identifying the coding regions in chromosomes is through the characterization of the nucleotide sequences of the detectable mRNAs (using the corresponding cDNAs). The mRNAs (and the cDNAs produced from them) lack introns, regulatory DNA sequences, and the nonessential "spacer" DNA that lies between genes. It is therefore useful to sequence large numbers of cDNAs to produce a very large database of the coding sequences of an organism. These sequences are then readily used to distinguish the exons from the introns in the long chromosomal DNA sequences that correspond to genes.

## The Genomes of Many Organisms Have Been Fully Sequenced

Owing in large part to the automation of DNA sequencing, the genomes of many organisms have been fully sequenced; these include plant chloroplasts and animal mitochondria, large numbers of bacteria, and archaea, and many of the model organisms that are studied routinely in the laboratory, including many yeasts, a nematode worm, the fruit fly *Drosophila*, the model plant *Arabidopsis*, the mouse, dog, chimpanzee, and, last but not least, humans. Researchers have also deduced the complete DNA sequences for a wide variety of human pathogens. These include the bacteria that cause cholera, tuberculosis, syphilis, gonorrhea, Lyme disease, and stomach ulcers, as well as hundreds of viruses—including smallpox virus and Epstein–Barr virus (which causes infectious mononucleosis). Examination of the genomes of these pathogens provides clues about what makes them virulent and will also point the way to new and more effective treatments.

*Haemophilus influenzae* (a bacterium that can cause ear infections and meningitis in children) was the first organism to have its complete genome sequence—all 1.8 million nucleotide pairs—determined by the *shotgun sequencing method*, the most common strategy used today. In the shotgun method, long sequences of DNA are broken apart randomly into many shorter fragments. Each fragment is then sequenced and a computer is used to order these pieces into a whole chromosome or genome, using sequence overlap to guide the assembly. The shotgun method is the technique of choice for

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 209 of 240

sequencing small genomes. Although larger, more repetitive genome sequences are more challenging to assemble, the shotgun method—in combination with the analysis of large DNA fragments cloned in BACs—has played a key role in their sequencing as well.

With new sequences appearing at a steadily accelerating pace in the scientific literature, comparison of the complete genome sequences of different organisms allows us to trace the evolutionary relationships among genes and organisms, and to discover genes and predict their functions (discussed in Chapters 3 and 4). Assigning functions to genes often involves comparing their sequences with related sequences from model organisms that have been well characterized in the laboratory, such as the bacterium *E. coli*, the yeasts *S. cerevisiae* and *S. pombe,* the nematode worm *C. elegans,* and the fruit fly *Drosophila* (discussed in Chapter 1).

Although the organisms whose genomes have been sequenced share many biochemical pathways and possess many proteins that are homologous in their amino acid sequence or structure, the functions of a very large number of newly identified proteins remain unknown. Depending on the organism, some 15–40% of the proteins encoded by a sequenced genome do not resemble any protein that has been studied biochemically. This observation underscores a limitation of the emerging field of genomics: although comparative analysis of genomes reveals a great deal of information about the relationships between genes and organisms, it often does not provide immediate information about how these genes function, or what roles they have in the physiology of an organism. Comparison of the full gene complement of several thermophilic bacteria, for example, does not reveal why these bacteria thrive at temperatures exceeding 70°C. And examination of the genome of the incredibly radioresistant bacterium *Deinococcus radiodurans* does not explain how this organism can survive a blast of radiation that can shatter glass. Further biochemical and genetic studies, like those described in the other sections of this chapter, are required to determine how genes, and the proteins they produce, function in the context of living organisms.

## Summary

*DNA cloning allows a copy of any specific part of a DNA or RNA sequence to be selected from the millions of other sequences in a cell and produced in unlimited amounts in pure form. DNA sequences can be amplified after cutting chromosomal DNA with a restriction nuclease and inserting the resulting DNA fragments into the chromosome of a self-replicating genetic element such as a virus or a plasmid. Plasmid vectors are generally used, and the resulting "genomic DNA library" is housed in millions of bacterial cells, each carrying a different cloned DNA fragment. Individual cells from this library that are allowed to proliferate produce large amounts of a single cloned DNA fragment. The polymerase chain reaction (PCR) allows DNA cloning to be performed directly with a thermostable DNA polymerase—provided that the DNA sequence of interest is already known.*

*The procedures used to obtain DNA clones that correspond in sequence to mRNA molecules are the same except that a DNA copy of the mRNA sequence, called cDNA, is first made. Unlike genomic DNA clones, cDNA clones lack intron sequences, making them the clones of choice for analyzing the protein product of a gene.*

*Nucleic acid hybridization reactions provide a sensitive means of detecting a gene or any other nucleotide sequence of interest. Under stringent hybridization conditions (a combination of solvent and temperature at which even a perfect double helix is barely stable), two strands can pair to form a "hybrid" helix only if their nucleotide sequences are almost perfectly complementary. The enormous specificity of this hybridization reaction allows any single-stranded sequence of nucleotides to be labeled with a radioisotope or chemical and used as a probe to find a complementary partner strand, even in a cell or cell extract that contains millions of different DNA and RNA sequences. Probes of this type are widely used to detect the nucleic acids corresponding to specific genes, both to facilitate their purification and characterization, and to localize them in cells, tissues, and organisms.*

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 210 of 240

*The nucleotide sequence of DNA can be determined rapidly and simply by using highly automated techniques based on the dideoxy method for sequencing DNA. This technique has made it possible to determine the complete DNA sequences of the genomes of many organisms. Comparison of the genome sequences of different organisms allows us to trace the evolutionary relationships among genes and organisms, and it has proved valuable for discovering new genes and predicting their functions.*

*Taken together, these techniques for analyzing and manipulating DNA have made it possible to identify, isolate, and sequence genes from any organism of interest. Related technologies allow scientists to produce the protein products of these genes in the large quantities needed for detailed analyses of their structure and function, as well as for medical purposes.*

## STUDYING GENE EXPRESSION AND FUNCTION

Ultimately, one wishes to determine how genes—and the proteins they encode—function in the intact organism. Although it may seem counterintuitive, one of the most direct ways to find out what a gene does is to see what happens to the organism when that gene is missing. Studying mutant organisms that have acquired changes or deletions in their nucleotide sequences is a time-honored practice in biology and forms the basis of the important field of **genetics**. Because mutations can disrupt cell processes, mutants often hold the key to understanding gene function. In the classical genetic approach, one begins by isolating mutants that have an interesting or unusual appearance: fruit flies with white eyes or curly wings, for example. Working backward from the **phenotype**—the appearance or behavior of the individual—one then determines the organism's **genotype**, the form of the gene responsible for that characteristic (**Panel 8–1**).

Today, with numerous genome sequences available, the exploration of gene function often begins with a DNA sequence. Here, the challenge is to translate sequence into function. One approach, discussed earlier in the chapter, is to search databases for well-characterized proteins that have similar amino acid sequences to the protein encoded by a new gene, and from there employ some of the methods described in the previous section to explore the gene's function further. But to determine directly a gene's function in a cell or organism, the most effective approach involves studying mutants that either lack the gene or express an altered version of it. Determining which cell processes have been disrupted or compromised in such mutants will usually shed light on a gene's biological role.

In this section, we describe several approaches to determining a gene's function, starting from a DNA sequence or an organism with an interesting phenotype. We begin with the classical genetic approach, which starts with a *genetic screen* for isolating mutants of interest and then proceeds toward identification of the gene or genes responsible for the observed phenotype. We then describe the set of techniques that are collectively called *reverse genetics*, in which one begins with a gene or gene sequence and attempts to determine its function. This approach often involves some intelligent guesswork—searching for homologous sequences and determining when and where a gene is expressed—as well as generating mutant organisms and characterizing their phenotype.

### Classical Genetics Begins by Disrupting a Cell Process by Random Mutagenesis

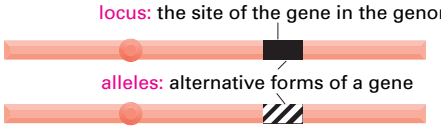
Before the advent of gene cloning technology, most genes were identified by the abnormalities produced when the gene was mutated. This classical genetic approach—identifying the genes responsible for mutant phenotypes—is most easily performed in organisms that reproduce rapidly and are amenable to genetic manipulation, such as bacteria, yeasts, nematode worms, and fruit flies. Although spontaneous mutants can sometimes be found by examining extremely large populations—thousands or tens of thousands of individual

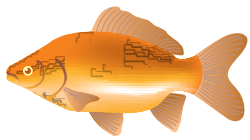
LCY Biotechnology Holding, Inc.
Ex. 1008
Page 211 of 240

## GENES AND PHENOTYPES

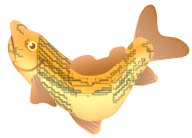
**Gene:** a functional unit of inheritance, usually corresponding to the segment of DNA coding for a single protein.

**Genome:** all of an organism's DNA sequences.



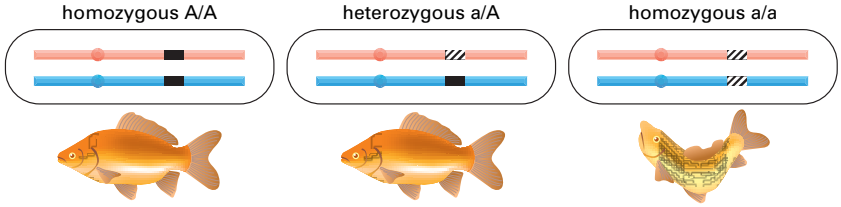
**Wild-type:** the normal, naturally occurring type



**Mutant:** differing from the wild-type because of a genetic change (a mutation)

**GENOTYPE:** the specific set of alleles forming the genome of an individual

**PHENOTYPE:** the visible character of the individual

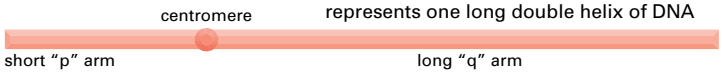


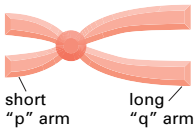

allele A is dominant (relative to a); allele a is recessive (relative to A)

In the example above, the phenotype of the heterozygote is the same as that of one of the homozygotes; in cases where it is different from both, the two alleles are said to be co-dominant.

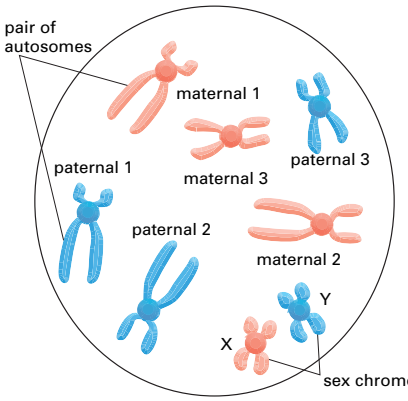
## CHROMOSOMES

a chromosome at the beginning of the cell cycle, in $G_1$ phase; the single long bar represents one long double helix of DNA



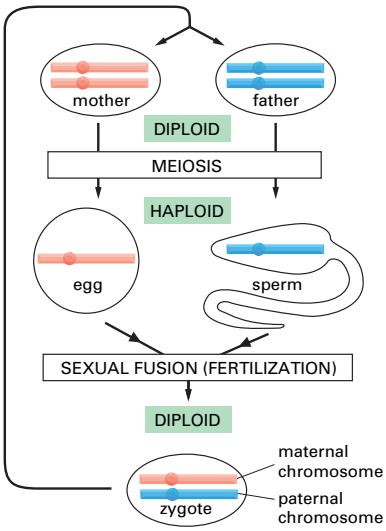




a chromosome near the end of the cell cycle, in metaphase; it is duplicated and condensed, consisting of two identical sister chromatids (each containing one DNA double helix) joined at the centromere.



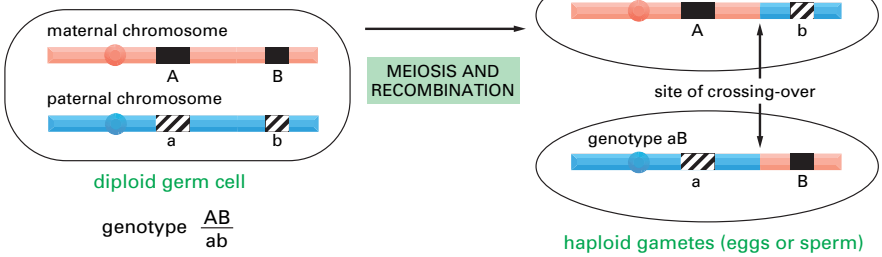

A normal diploid chromosome set, as seen in a metaphase spread, prepared by bursting open a cell at metaphase and staining the scattered chromosomes. In the example shown schematically here, there are three pairs of autosomes (chromosomes inherited symmetrically from both parents, regardless of sex) and two sex chromosomes—an X from the mother and a Y from the father. The numbers and types of sex chromosomes and their role in sex determination are variable from one class of organisms to another, as is the number of pairs of autosomes.

## THE HAPLOID–DIPLOID CYCLE OF SEXUAL REPRODUCTION




For simplicity, the cycle is shown for only one chromosome/chromosome pair.

## MEIOSIS AND GENETIC RECOMBINATION



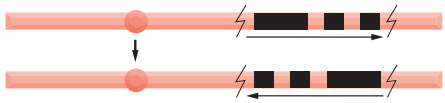

The greater the distance between two loci on a single chromosome, the greater is the chance that they will be separated by crossing over occurring at a site between them. If two genes are thus reassorted in x% of gametes, they are said to be separated on a chromosome by a genetic map distance of x map units (or x centimorgans).

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 212 of 240

## TYPES OF MUTATIONS



POINT MUTATION: maps to a single site in the genome, corresponding to a single nucleotide pair or a very small part of a single gene

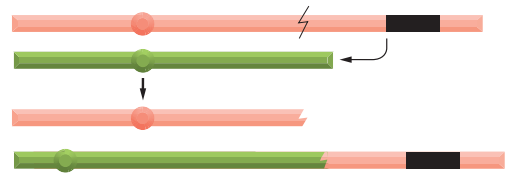


INVERSION: inverts a segment of a chromosome

lethal mutation: causes the developing organism to die prematurely.
conditional mutation: produces its phenotypic effect only under certain conditions, called the *restrictive* conditions. Under other conditions—the *permissive* conditions—the effect is not seen. For a *temperature-sensitive* mutation, the restrictive condition typically is high temperature, while the permissive condition is low temperature.
loss-of-function mutation: either reduces or abolishes the activity of the gene. These are the most common class of mutations. Loss-of-function mutations are usually *recessive*—the organism can usually function normally as long as it retains at least one normal copy of the affected gene.
null mutation: a loss-of-function mutation that completely abolishes the activity of the gene.



DELETION: deletes a segment of a chromosome



TRANSLOCATION: breaks off a segment from one chromosome and attaches it to another

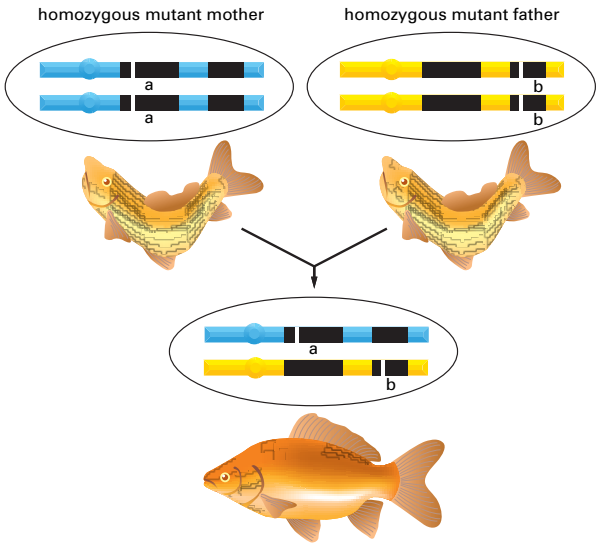
gain-of-function mutation: increases the activity of the gene or makes it active in inappropriate circumstances; these mutations are usually *dominant.*
dominant-negative mutation: dominant-acting mutation that blocks gene activity, causing a loss-of-function phenotype even in the presence of a normal copy of the gene. This phenomenon occurs when the mutant gene product interferes with the function of the normal gene product.
suppressor mutation: suppresses the phenotypic effect of another mutation, so that the double mutant seems normal. An *intragenic* suppressor mutation lies within the gene affected by the first mutation; an *extragenic* suppressor mutation lies in a second gene—often one whose product interacts directly with the product of the first.

## TWO GENES OR ONE?

Given two mutations that produce the same phenotype, how can we tell whether they are mutations in the same gene? If the mutations are recessive (as they most often are), the answer can be found by a complementation test.

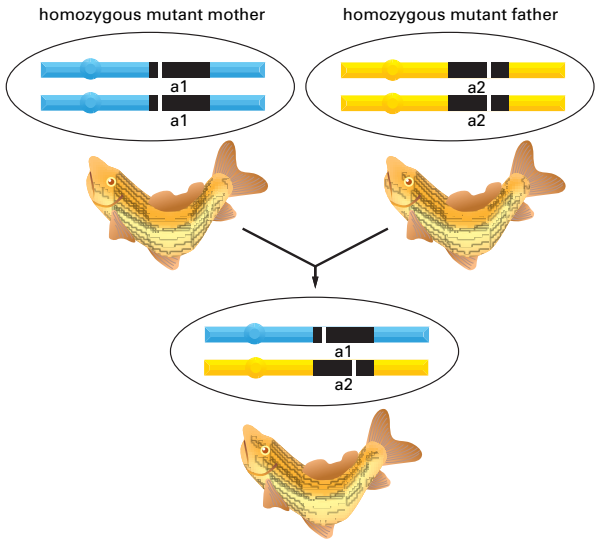
In the simplest type of complementation test, an individual who is homozygous for one mutation is mated with an individual who is homozygous for the other. The phenotype of the offspring gives the answer to the question.

### COMPLEMENTATION: MUTATIONS IN TWO DIFFERENT GENES




hybrid offspring shows normal phenotype: one normal copy of each gene is present

### NONCOMPLEMENTATION: TWO INDEPENDENT MUTATIONS IN THE SAME GENE




hybrid offspring shows mutant phenotype: no normal copies of the mutated gene are present

organisms—isolating mutant individuals is much more efficient if one generates mutations with chemicals or radiation that damage DNA. By treating organisms with such mutagens, very large numbers of mutant individuals can be created quickly and then screened for a particular defect of interest, as we discuss shortly.

An alternative approach to chemical or radiation mutagenesis is called *insertional mutagenesis*. This method relies on the fact that exogenous DNA inserted randomly into the genome can produce mutations if the inserted fragment interrupts a gene or its regulatory sequences. The inserted DNA, whose sequence is known, then serves as a molecular tag that aids in the subsequent identification and cloning of the disrupted gene (**Figure 8–53**). In *Drosophila*, the use of the transposable P element to inactivate genes has revolutionized the study of gene function in the fly. Transposable elements (see Table 5–3, p. 318) have also been used to generate mutations in bacteria, yeast, mice, and the flowering plant *Arabidopsis*.

Such classical genetic studies are well suited for dissecting biological processes in experimental organisms, but how can we study gene function in humans? Unlike the genetically accessible organisms we have been discussing, humans do not reproduce rapidly, and they cannot be intentionally treated with mutagens. Moreover, any human with a serious defect in an essential process, such as DNA replication, would die long before birth.

There are two main ways that we can study human genes. First, because genes and gene functions have been so highly conserved throughout evolution, the study of less complex model organisms reveals critical information about similar genes and processes in humans. The corresponding human genes can then be studied further in cultured human cells. Second, many mutations that are not lethal—tissue-specific defects in lysosomes or cell-surface receptors, for example—have arisen spontaneously in the human population. Analyses of the phenotypes of the affected individuals, together with studies of their cultured cells, have provided many unique insights into important human cell functions. Although such mutations are rare, they are very efficiently discovered because of a unique human property: the mutant individuals call attention to themselves by seeking special medical care.



**Figure 8–53 Insertional mutant of the snapdragon, Antirrhinum.** A mutation in a single gene coding for a regulatory protein causes leafy shoots to develop in place of flowers. The mutation allows cells to adopt a character that would be appropriate to a different part of the normal plant. The mutant plant is on the *left,* the normal plant on the *right*. (Courtesy of Enrico Coen and Rosemary Carpenter.)

## Genetic Screens Identify Mutants with Specific Abnormalities

Once a collection of mutants in a model organism such as yeast or fly has been produced, one generally must examine thousands of individuals to find the altered phenotype of interest. Such a search is called a **genetic screen**, and the larger the genome, the less likely it is that any particular gene will be mutated. Therefore, the larger the genome of an organism, the bigger the screening task becomes. The phenotype being screened for can be simple or complex. Simple phenotypes are easiest to detect: one can screen many organisms rapidly, for example, for mutations that make it impossible for the organism to survive in the absence of a particular amino acid or nutrient.

More complex phenotypes, such as defects in learning or behavior, may require more elaborate screens (**Figure 8–54**). But even genetic screens that are used to dissect complex physiological systems should be as simple as possible in design, and, if possible, should permit the simultaneous examination of large numbers of mutants. As an example, one particularly elegant screen was designed to search for genes involved in visual processing in zebrafish. The basis of this screen, which monitors the fishes' response to motion, is a change in behavior. Wild-type fish tend to swim in the direction of a perceived motion, whereas mutants with defects in their visual processing systems swim in random directions—a behavior that is easily detected. One mutant discovered in this screen is called *lakritz*, which is missing 80% of the retinal ganglion cells that help to relay visual signals from the eye to the brain. As the cellular organization of the zebrafish retina is similar to that of all vertebrates, the study of such mutants should also provide insights into visual processing in humans.

Because defects in genes that are required for fundamental cell processes—RNA synthesis and processing or cell-cycle control, for example—are usually

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 214 of 240




**Figure 8–54 A behavioral phenotype detected in a genetic screen.** (A) Wild-type *C. elegans* engage in social feeding. The worms migrate around until they encounter their neighbors and commence feeding on bacteria. (B) Mutant animals feed by themselves. (Courtesy of Cornelia Bargmann, *Cell* 94: cover, 1998. With permission from Elsevier.)

lethal, the functions of these genes are often studied in individuals with **conditional mutations**. The mutant individuals function normally as long as "permissive" conditions prevail, but demonstrate abnormal gene function when subjected to "nonpermissive" (restrictive) conditions. In organisms with *temperature-sensitive mutations,* for example, the abnormality can be switched on and off experimentally simply by changing the temperature; thus, a cell containing a temperature-sensitive mutation in a gene essential for survival will die at a nonpermissive temperature but proliferate normally at the permissive temperature (**Figure 8–55**). The temperature-sensitive gene in such a mutant usually contains a point mutation that causes a subtle change in its protein product.

Many temperature-sensitive mutations were found in the bacterial genes that encode the proteins required for DNA replication. The mutants were identified by screening populations of mutagen-treated bacteria for cells that stop making DNA when they are warmed from 30°C to 42°C. These mutants were later used to identify and characterize the corresponding DNA replication proteins (discussed in Chapter 5). Similarly, screens for temperature-sensitive mutations led to the identification of many proteins involved in regulating the cell cycle, as well as many proteins involved in moving proteins through the secretory pathway in yeast (see Panel 13–1). Related screening approaches demonstrated the function of enzymes involved in the principal metabolic pathways of bacteria and yeast (discussed in Chapter 2) and identified many of the gene products responsible for the orderly development of the *Drosophila* embryo (discussed in Chapter 22).

## Mutations Can Cause Loss or Gain of Protein Function

Gene mutations are generally classed as "loss of function" or "gain of function." A loss of function mutation results in a gene product that either does not work or works too little; thus, it reveals the normal function of the gene. A gain of function mutation results in a gene product that works too much, works at the wrong time or place, or works in a new way (**Figure 8–56**).

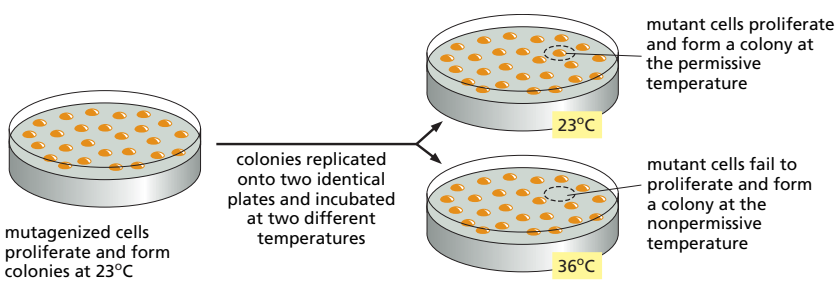



**Figure 8–55 Screening for temperature-sensitive bacterial or yeast mutants.** Mutagenized cells are plated out at the permissive temperature. They divide and form colonies, which are transferred to two identical Petri dishes by replica plating. One of these plates is incubated at the permissive temperature, the other at the nonpermissive temperature. Cells containing a temperature-sensitive mutation in a gene essential for proliferation can divide at the normal, permissive temperature but fail to divide at the elevated, nonpermissive temperature.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 215 of 240

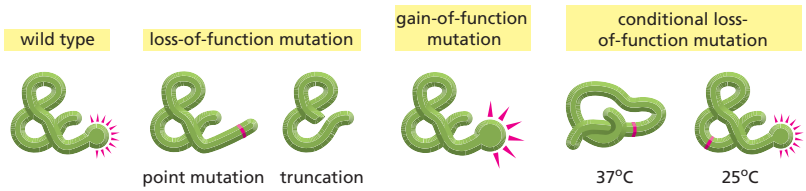



**Figure 8–56 Gene mutations that affect their protein product in different ways.** In this example, the wild-type protein has a specific cell function denoted by the *red rays.* Mutations that eliminate this function, increase the function, or render the function sensitive to higher temperatures are shown. The temperature-sensitive conditional mutant protein carries an amino acid substitution *(red)* that prevents its proper folding at 37°C, but allows the protein to fold and function normally at 25°C. Such conditional mutations are especially useful for studying essential genes; the organism can be grown under the permissive condition and then moved to the nonpermissive condition to study the function of the gene.

An important early step in the genetic analysis of any mutant cell or organism is to determine whether the mutation causes a loss or a gain of function. A standard test is to determine whether the mutation is *dominant* or *recessive.* A dominant mutation is one that still causes the mutant phenotype, in the presence of a single copy of the wild-type gene. A recessive mutation is one that is no longer able to cause the mutant phenotype in the presence of a single wild-type copy of the gene. Although cases have been described in which a loss-of-function mutation is dominant or a gain-of-function mutation is recessive, in the vast majority of cases, recessive mutations are loss of function, and dominant mutations are gain of function. It is easy to determine if a mutation is dominant or recessive. One simply mates a mutant with a wild-type to obtain diploid cells or organisms. The progeny from the mating will be heterozygous for the mutation. If the mutant phenotype is no longer observed, one can conclude that the mutation is recessive and is very likely to be a loss-of-function mutation.

## Complementation Tests Reveal Whether Two Mutations Are in the Same Gene or Different Genes

A large-scale genetic screen can turn up many different mutations that show the same phenotype. These defects might lie in different genes that function in the same process, or they might represent different mutations in the same gene. Alternative forms of a gene are known as **alleles**. The most common difference between alleles is a substitution of a single nucleotide pair, but different alleles can also bear deletions, substitutions, and duplications. How can we tell, then, whether two mutations that produce the same phenotype occur in the same gene or in different genes? If the mutations are recessive—if, for example, they represent a loss of function of a particular gene—a **complementation test** can be used to ascertain whether the mutations fall in the same or in different genes. To test complementation in a diploid organism, an individual that is homozygous for one mutation—that is, it possesses two identical alleles of the mutant gene in question—is mated with an individual that is homozygous for the other mutation. If the two mutations are in the same gene, the offspring show the mutant phenotype, because they still will have no normal copies of the gene in question (see Panel 8–1). If, in contrast, the mutations fall in different genes, the resulting offspring show a normal phenotype, because they retain one normal copy (and one mutant copy) of each gene; the mutations thereby complement one another and restore a normal phenotype. Complementation testing of mutants identified during genetic screens has revealed, for example, that 5 genes are required for yeast to digest the sugar galactose, 20 genes are needed for *E. coli* to build a functional flagellum, 48 genes are involved in assembling bacteriophage T4 viral particles, and hundreds of genes are involved in the development of an adult nematode worm from a fertilized egg.

## Genes Can Be Ordered in Pathways by Epistasis Analysis

Once a set of genes involved in a particular biological process has been identified, the next step is often to determine in which order the genes function. Gene order is perhaps easiest to explain for metabolic pathways, where, for example, enzyme A is necessary to produce the substrate for enzyme B. In this case, we would say that the gene encoding enzyme A acts before (upstream of) the gene encoding enzyme B in the pathway. Similarly, where one protein regulates the

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 216 of 240

activity of another protein, we would say that the former gene acts before the latter. Gene order can, in many cases, be determined purely by genetic analysis without any knowledge of the mechanism of action of the gene products involved.

Suppose we have a biosynthetic process consisting of a sequence of steps, such that performance of a step B is conditional on completion of the preceding step A; and suppose *gene A* is required for step A, and *gene B* is required for step B. Then a null mutation (a mutation that abolishes function) in gene *A* will arrest the process at step A, regardless of whether gene *B* is functional or not, whereas a null mutation in gene *B* will cause arrest at step B only if gene *A* is still active. In such a case, gene *A* is said to be *epistatic* to gene *B*. By comparing the phenotypes of the different combinations of mutations, we can therefore discover the order in which the genes act. This type of analysis is called **epistasis analysis**. As an example, the pathway of protein secretion in yeast has been analyzed in this way. Different mutations in this pathway cause proteins to accumulate aberrantly in the endoplasmic reticulum (ER) or in the Golgi apparatus. When a yeast cell is engineered to carry both a mutation that blocks protein processing in the ER *and* a mutation that blocks processing in the Golgi apparatus, proteins accumulate in the ER. This indicates that proteins must pass through the ER before being sent to the Golgi before secretion (**Figure 8–57**). Strictly speaking, an epistasis analysis can only provide information about gene order in a pathway when both mutations are null alleles. When the mutations retain partial function, their epistasis interactions can be difficult to interpret.

Sometimes, a double mutant will show a new or more severe phenotype than either single mutant alone. This type of genetic interaction is called a *synthetic* phenotype, and if the phenotype is death of the organism, it is called *synthetic lethality.* In most cases, a synthetic phenotype indicates that the two genes act in two different parallel pathways, either of which is capable of mediating the same cell process. Thus, when both pathways are disrupted in the double mutant, the process fails altogether, and the synthetic phenotype is observed.

## Genes Identified by Mutations Can Be Cloned

Once the mutant organisms are produced in a genetic screen, the next task is identifying the gene or genes responsible for the altered phenotype. If the phenotype has been produced by insertional mutagenesis, locating the disrupted gene is fairly simple. DNA fragments containing the insertion (a transposon or a retrovirus, for example) are collected and amplified by PCR, and the nucleotide sequence of the flanking DNA is determined. Genome databases can then be searched for open reading frames containing this flanking sequence.

If a DNA-damaging chemical was used to generate the mutations, identifying the inactivated gene is often more laborious, but it can be accomplished by several different approaches. In one, the first step is to experimentally determine the gene's location in the genome. To map a newly discovered gene, its rough chromosomal location is first determined by assessing how far the gene lies from other known genes in the genome. Estimating the distance between genetic loci is usually done by *linkage analysis*, a technique that relies on the tendency for genes that lie near one another on a chromosome to be inherited together. Even closely linked genes, however, can be separated by

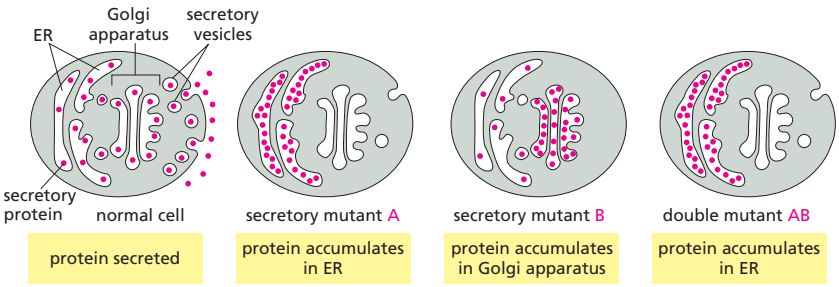



**Figure 8–57 Using genetics to determine the order of function of genes.** In normal cells, secretory proteins are loaded into vesicles, which fuse with the plasma membrane to secrete their contents into the extracellular medium. Two mutants, A and B, fail to secrete proteins. In mutant A, secretory proteins accumulate in the ER. In mutant B, secretory proteins accumulate in the Golgi. In the double mutant AB, proteins accumulate in the ER; this indicates that the gene defective in mutant A acts before the gene defective in mutant B in the secretory pathway.

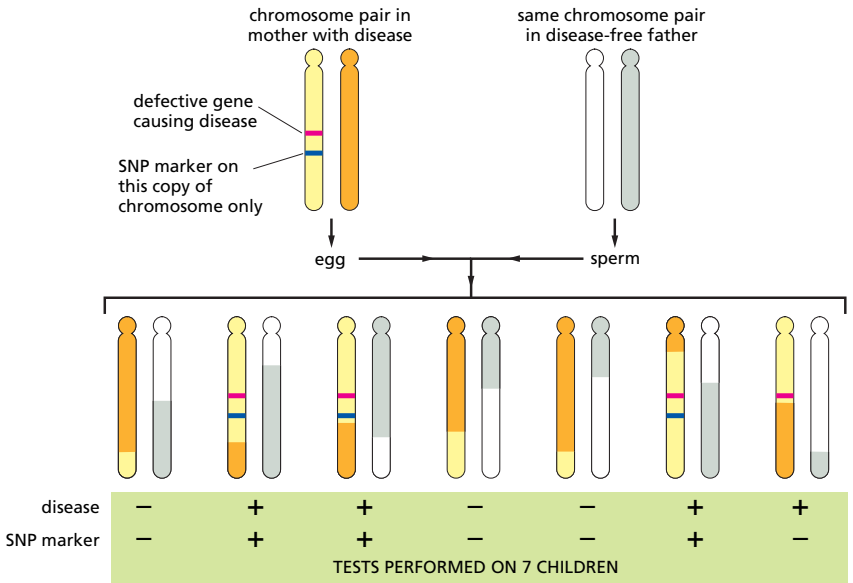
LCY Biotechnology Holding, Inc.
Ex. 1008
Page 217 of 240

recombination during meiosis. The larger the distance between two genetic loci, the greater the chance that they will be separated by a crossover (see Panel 8–1). By calculating the recombination frequency between two genes, the approximate distance between them can be determined. If the position of one gene in the genome is known, that of the second gene can thereby be estimated.

Because genes are not always located close enough to one another to allow a precise pinpointing of their position, linkage analyses often rely on physical markers along the genome for estimating the location of an unknown gene. These markers are generally short stretches of nucleotides, with a known sequence and genome location, that can exist in at least two allelic forms. The simplest markers are *single-nucleotide polymorphisms (SNPs),* short sequences that differ by one nucleotide pair among individuals in a population. SNPs can be detected by hybridization techniques. Many such physical markers, distributed all along the length of chromosomes, have been collected for a variety of organisms. If the distribution of these markers is sufficiently dense, one can, through a linkage analysis that tests for the tight co-inheritance of one or more SNPs with the mutant phenotype, narrow the potential location of a gene to a chromosomal region that may contain only a few gene sequences. These are then considered candidate genes, and their structure and function can be tested directly to determine which gene is responsible for the original mutant phenotype.

## Human Genetics Presents Special Problems and Opportunities

Although genetic experimentation on humans is considered unethical and is legally banned, humans do suffer from a large variety of genetic disorders. The linkage analysis described above can be used to identify the genes responsible for these heritable conditions. Such studies require DNA samples from a large number of families affected by the disease. These samples are examined for the presence of physical markers such as SNPs that seem to be closely linked to the disease gene, in that they are always inherited by individuals who have the disease and not by their unaffected relatives. The disease gene is then located as described above (**Figure 8–58**). The genes for cystic fibrosis and Huntington's disease, for example, were discovered in this way.




CONCLUSION: gene causing disease is co-inherited with SNP marker from diseased mother in 75% of the diseased progeny. If this same correlation is observed in other families that have been examined, the gene causing disease is mapped to this chromosome close to the SNP. Note that an SNP that is either far away from the gene on the same chromosome or located on a different chromosome from the gene of interest will be co-inherited only 50% of the time.

**Figure 8–58 Genetic linkage analysis using physical markers on DNA to find a human gene.** In this example, the co-inheritance of a specific human phenotype (here a genetic disease) with an SNP marker. If individuals who inherit the disease nearly always inherit a particular SNP marker, then the gene causing the disease and the SNP are likely to be close together on the chromosome, as shown here. To prove that an observed linkage is statistically significant, hundreds of individuals may need to be examined. Note that the linkage will not be absolute unless the SNP marker is located in the gene itself. Thus, occasionally the SNP will be separated from the disease gene by crossing over during meiosis in the formation of the egg or sperm: this has happened in the case of the chromosome pair on the far right. When working with a sequenced genome, this procedure would be repeated with SNPs located on either side of the initial SNP, until a 100% co-inheritance is found.

Note that the egg and sperm will each contribute only one chromosome of each pair from the parent to the child.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 218 of 240

## Human Genes Are Inherited in Haplotype Blocks, Which Can Aid in the Search for Mutations That Cause Disease

With the complete human genome sequence in hand, we can now study human genetics in a way that was impossible only a few years ago. For example, we can begin to identify those DNA differences that distinguish one individual from another. No two humans (with the exception of identical twins) have the same genome. Each of us carries a set of polymorphisms—differences in nucleotide sequence—that make us unique. These polymorphisms can be used as markers for building genetic maps and performing genetic analyses to link particular polymorphisms with specific diseases or predispositions to disease.

The problem is that any two humans typically differ by about 0.1% in their nucleotide sequences (approximately one nucleotide difference every 1000 nucleotides). This translates to about 3 million differences between one person and another. Theoretically, one would need to search through all 3 million of those polymorphisms to identify the one or two that are responsible for a particular heritable disease or disease predisposition. To reduce the number of polymorphisms we need to examine, researchers are taking advantage of the recent discovery that human genes tend to be inherited in blocks.

The human species is relatively young, and it is thought that we are descended from a relatively small population of individuals who lived in Africa about 100,000 years ago. Because only a few thousand generations separate us from this ancestral population, large segments of human chromosomes have passed from parent to child unaltered by the recombination events that occur in meiosis. In fact, we observe that certain sets of alleles (including SNPs) are inherited in large blocks within chromosomes. These ancestral chromosome segments—sets of alleles that have been inherited in clusters with little genetic rearrangement across the generations—are called **haplotype blocks**. Like genes, SNPs, and other genetic markers—which exist in different alleleic forms—haplotype blocks also come in a limited number of "flavors" that are common in the human population, each of which represents an allele combination passed down from a shared ancestor long ago.

Researchers are now constructing a human genome map based on these haplotype blocks—called a **haplotype map** (**hapmap**). Geneticists hope that the human haplotype map will make the search for disease-causing and disease-susceptibility genes a much more manageable task. Instead of searching through each of the many millions of SNPs in the human population, one need only search through a considerably smaller set of selected SNPs to identify the haplotype block that appears to be inherited by individuals with the disease. (These searches still involve DNA samples from large numbers of people, and SNPs are now typically scored using robotic technologies.) If a specific haplotype block is more common among people with the disease than in unaffected individuals, the mutation linked to that disease will likely be located in that same segment of DNA (**Figure 8–59**). Researchers can then zero in on the specific region within the block to search for the specific gene associated with the disease. This approach should, in principle, allow one to analyze the genetics of those common diseases in which multiple genes confer susceptibility.

A detailed examination of haplotype blocks can even tell us whether a particular allele has been favored by natural selection. As a rule, when a new allele of a gene arises that does not confer a selective advantage on the individual, it will take a long time for that allele to become common in the population. The more common—and therefore older—such an allele is, the smaller should be the haplotype block that surrounds it, because it will have had many chances of being separated from its neighboring variations by the recombination events that occur in meiosis generation after generation.

A new allele may quickly spread in a population, however, if it confers some dramatic advantage on the organism. For example, mutations or variations that make an organism more resistant to an infection might be selected for because organisms with this variation would be more likely to survive and pass the

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 219 of 240

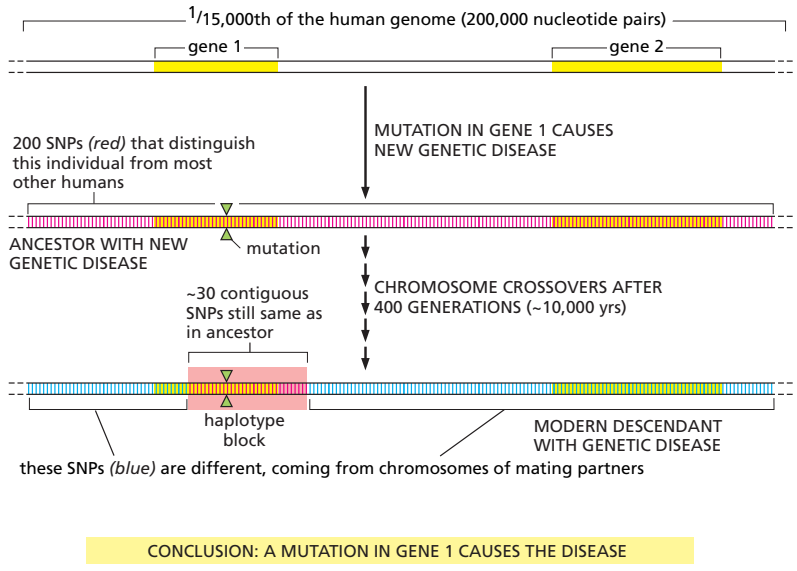



**Figure 8–59 Tracing the inheritance of SNPs within haplotype blocks to reveal the location of a disease-causing gene.** An ancestor who acquires a disease-causing mutation in gene 1 will pass that mutation along to his or her descendants. Part of this gene is embedded within a haplotype block *(red shading)*—a cluster of variations (about 30 SNPs) that have been passed along from the ancestor in a continuous chunk. In the 400 generations that separate the ancestor from modern descendants with the disease, SNPs located over most of the ancestral 200,000-nucleotide-pair region shown have been shuffled by meiotic recombination in the descendant genome *(blue)*. (Note that the overlap of yellow and red is seen as *orange,* and the overlap of yellow and blue is seen as *green.*) The 30 SNPs within the haplotype block, however, have been inherited as a group, as no crossover events have yet separated them. To locate a gene that causes the inherited disease, the SNP patterns in a number of people who have the disease need to be analyzed. An individual with the disease will retain the ancestral pattern of SNPs located within the haplotype block shown, revealing that the disease-causing mutation is likely to lie within that haplotype block—thus in gene 1. The beauty of using haplotype maps for this type of linkage analysis is that only a fraction of the total SNPs need to be examined: one should be able to locate genes after searching through only about 10% of the 3 million useful SNPs present in the human genome.

mutation on to their offspring. Working with haplotype maps of individual genes, researchers have detected such positive selection for two human genes that confer resistance to malaria. The alleles that confer resistance are widespread in the population, but they are embedded in unusually large haplotype blocks, suggesting that they rose to prominence recently in the human gene pool (**Figure 8–60**).

In revealing the paths along which humans evolved, the human haplotype map provides a new window into our past; in helping us discover the genes that make us susceptible or resistant to disease, the map may also provide a rough guide to our individual futures.

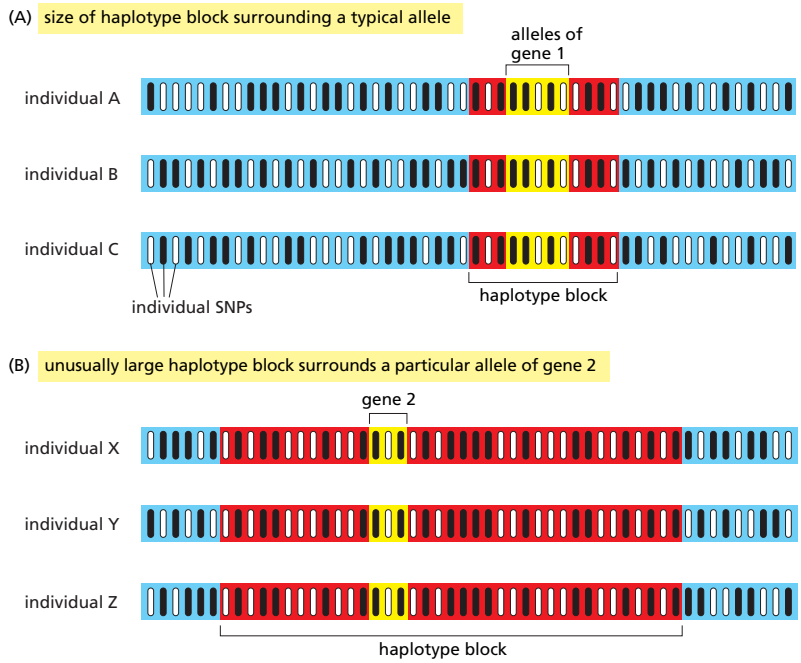



**Figure 8–60 Identification of alleles that have been selected for in fairly recent human history by the unusually large haplotype blocks in which they are embedded.** The SNPs are indicated in this diagram by vertical bars, which are shown as *white* or *black* according to their DNA sequence. Haplotype blocks are shaded in *red,* genes in *yellow,* and the rest of the chromosome in *blue.* These data suggest that this particular allele of gene 2 arose relatively recently in human history.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 220 of 240

## Complex Traits Are Influenced by Multiple Genes

A concert pianist might have an aunt who plays the violin. In another family, the parents and the children might all be fat. In a third family, the grandmother might be an alcoholic, and her grandson might abuse drugs. To what extent are such characteristics—musical ability, obesity, and addiction—inherited genetically? This is a very difficult question to answer. Some traits or diseases "run in families" but appear in only a few relatives or with no easily discernible pattern.

Characteristics that do not follow simple (sometimes called Mendelian) patterns of inheritance but have a genetically inherited component are termed **complex traits**. These traits are often **polygenic**; that is, they are influenced by multiple genes, each of which makes a small contribution to the phenotype in question. The effects of these genes are additive, which means that, together, they produce a continuum of varying features within the population. Individually, the genes that contribute to a polygenic trait are distributed to offspring in simple patterns, but because they all influence the phenotype, the pattern of traits inherited by offspring is often highly complex.

A simple example of a polygenic trait is eye color, which is determined by enzymes that control the distribution and production of the pigment melanin: the more melanin produced, the darker the eye color. Because numerous genes contribute to the formation of melanin, eye color in humans shows enormous variation, from the palest gray to a dark chocolate brown.

Although diseases based on mutations in single genes (for example, sickle-cell anemia and hemophilia) were some of the earliest recognized human inherited phenotypes, only a small fraction of human traits are dictated by single genes. The most obvious human phenotypes—from height, weight, eye color, and hair color to intelligence, temperament, sociability, and humor—arise from the interaction of many genes. Multiple genes also almost certainly underlie a propensity for the most common human diseases: diabetes, heart disease, high blood pressure, allergies, asthma, and various mental illnesses, including major depression and schizophrenia. Researchers are exploring new strategies—including the use of the haplotype maps discussed earlier—to understand the complex interplay between genes that act together to determine many of our most "human" traits.

## Reverse Genetics Begins with a Known Gene and Determines Which Cell Processes Require Its Function

As we have seen, classical genetics starts with a mutant phenotype (or, in the case of humans, a range of characteristics) and identifies the mutations (and consequently the genes) responsible for it. Recombinant DNA technology, in combination with genome sequencing, has made possible a different type of genetic approach. Instead of beginning with a mutant organism and using it to identify a gene and its protein, an investigator can start with a particular gene and proceed to make mutations in it, creating mutant cells or organisms so as to analyze the gene's function. Because this approach reverses the traditional direction of genetic discovery—proceeding from genes to mutations, rather than vice versa—it is commonly referred to as **reverse genetics**.

Reverse genetics begins with a cloned gene, a protein with interesting properties that has been isolated from a cell, or simply a genome sequence. If the starting point is a protein, the gene encoding it is first identified and, if necessary, its nucleotide sequence is determined. The gene sequence can then be altered *in vitro* to create a mutant version. This engineered mutant gene, together with an appropriate regulatory region, is transferred into a cell where it can integrate into a chromosome, becoming a permanent part of the cell's genome. All of the descendants of the modified cell will now contain the mutant gene.

If the original cell used for the gene transfer is a fertilized egg, whole multicellular organisms can be obtained that contain the mutant gene, provided that the mutation does not cause lethality. In some of these animals, the altered gene

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 221 of 240

will be incorporated into the germ cells—a *germ-line mutation*—allowing the mutant gene to be passed on to their progeny.

## Genes Can Be Engineered in Several Ways

We have seen that mutant organisms lacking a particular gene may quickly reveal the function of the protein it encodes. For this reason, a gene "knockout"—in which both copies of the gene in a diploid organism have been inactivated or deleted—is a particularly useful type of mutation. However, there are many more types of genetic alterations available to the experimenter. For example, by altering the regulatory region of a gene before it is reintegrated into the genome, one can create mutant organisms in which the gene product is expressed at abnormally high levels, in the wrong tissue, or at the wrong time in development (**Figure 8–61**). By placing the gene under the control of an *inducible promoter*, the gene can be switched on or off at any time, and the effects observed. Inducible promoters that function in only a specific tissue can be used to monitor the effects of shutting the gene off (or turning it on) in that particular tissue. Finally, *dominant-negative* mutations are often employed particularly in those organisms in which it is simpler to add an altered gene to the genome than to replace the endogenous genes with it. The dominant-negative strategy exploits the fact that most proteins function as parts of larger protein complexes. The inclusion of just one nonfunctional component can often inactivate such complexes. Therefore, by designing a gene that produces large quantities of a mutant protein that is inactive but still able to assemble into the complex, it is often possible to produce a cell in which all the complexes are inactivated despite the presence of the normal protein (**Figure 8–62**).

As noted in the earlier discussion of classical genetics, if a protein is required for the survival of the cell (or the organism), a dominant-negative mutant will be inviable, making it impossible to test the function of the protein. To avoid this problem in reverse genetics, one can couple the mutant gene to an inducible promoter in order to produce the faulty gene product only on command—for example, in response to an increase in temperature or to the presence of a specific signal molecule.

In studying the action of a gene and the protein it encodes, one does not always wish to make drastic changes—flooding cells with huge quantities of the protein or eliminating a gene product entirely. It is sometimes useful to make slight changes in a protein's structure so that one can begin to dissect which portions of a protein are important for its function. The activity of an enzyme, for example, can be studied by changing a single amino acid in its active site. Special techniques are required to alter genes (and thus their protein products) in such subtle ways. The first step is often the chemical synthesis of a short DNA molecule containing the desired altered portion of the gene's nucleotide sequence. This synthetic DNA oligonucleotide is hybridized with single-stranded plasmid DNA that contains the DNA sequence to be altered, using conditions that allow imperfectly matched DNA strands to pair. The synthetic oligonucleotide will now serve as a primer for DNA synthesis by DNA polymerase, thereby generating a DNA double helix that incorporates the altered



**Figure 8–61 Ectopic misexpression of Wnt, a signaling protein that affects development of the body axis in the early *Xenopus* embryo.** In this experiment, mRNA coding for Wnt was injected into the ventral vegetal blastomere, inducing a second body axis (discussed in Chapter 22). (From S. Sokol et al., *Cell* 67:741–752, 1991. With permission from Elsevier.)

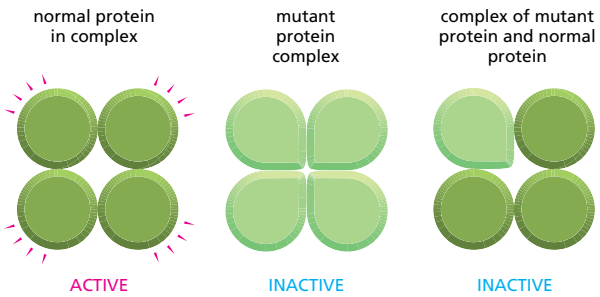



**Figure 8–62 A dominant-negative effect of a protein.** Here, a gene is engineered to produce a mutant protein that prevents the normal copies of the same protein from performing their function. In this simple example, the normal protein must form a multisubunit complex to be active, and the mutant protein blocks function by forming a mixed complex that is inactive. In this way, a single copy of a mutant gene located anywhere in the genome can inactivate the normal products produced by other gene copies.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 222 of 240

sequence into one of its two strands. After transfection, plasmids that carry the fully modified gene sequence are obtained. The modified DNA is then inserted into an expression vector so that the redesigned protein can be produced in the appropriate type of cells for detailed studies of its function. By changing selected amino acids in a protein in this way—a technique called **site-directed mutagenesis**—one can determine exactly which parts of the polypeptide chain are important for such processes as protein folding, interactions with other proteins, and enzymatic catalysis (**Figure 8–63**).

## Engineered Genes Can Be Inserted into the Germ Line of Many Organisms

Altered genes can be introduced into cells in a variety of ways. DNA can be microinjected into mammalian cells with a glass micropipette or introduced by a virus that has been engineered to carry foreign genes. In plant cells, genes are frequently introduced by a technique called particle bombardment: DNA samples are painted onto tiny gold beads and then literally shot through the cell wall with a specially modified gun. *Electroporation* is the method of choice for introducing DNA into bacteria and some other cells. In this technique, a brief electric shock renders the cell membrane temporarily permeable, allowing foreign DNA to enter the cytoplasm.

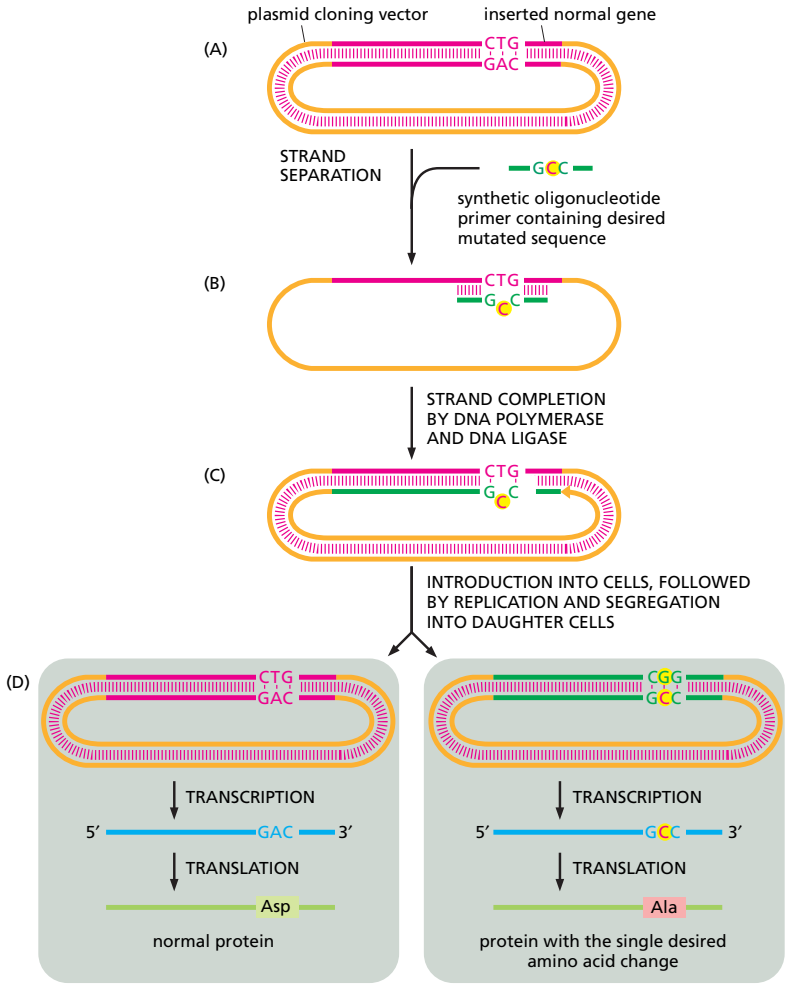



**Figure 8–63 The use of a synthetic oligonucleotide to modify the protein-coding region of a gene by site-directed mutagenesis.** (A) A recombinant plasmid containing a gene insert is separated into its two DNA strands. A synthetic oligonucleotide primer corresponding to part of the gene sequence but containing a single altered nucleotide at a predetermined point is added to the single-stranded DNA under conditions that permit imperfect DNA hybridization (see Figure 8–36). (B) The primer hybridizes to the DNA, forming a single mismatched nucleotide pair. (C) The recombinant plasmid is made double-stranded by *in vitro* DNA synthesis (starting from the primer) followed by sealing by DNA ligase. (D) The double-stranded DNA is introduced into a cell, where it is replicated. Replication using one strand of the template produces a normal DNA molecule, but replication using the other strand (the one that contains the primer) produces a DNA molecule carrying the desired mutation. Only half of the progeny cells will end up with a plasmid that contains the desired mutant gene. However, a progeny cell that contains the mutated gene can be identified, separated from other cells, and cultured to produce a pure population of cells, all of which carry the mutated gene. Only one of the many changes that can be engineered in this way is shown here. With an oligonucleotide of the appropriate sequence, more than one amino acid substitution can be made at a time, or one or more amino acids can be inserted or deleted. Although not shown in this figure, it is also possible to create a site-directed mutation by using the appropriate oligonucleotides and PCR (instead of plasmid replication) to amplify the mutated gene.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 223 of 240

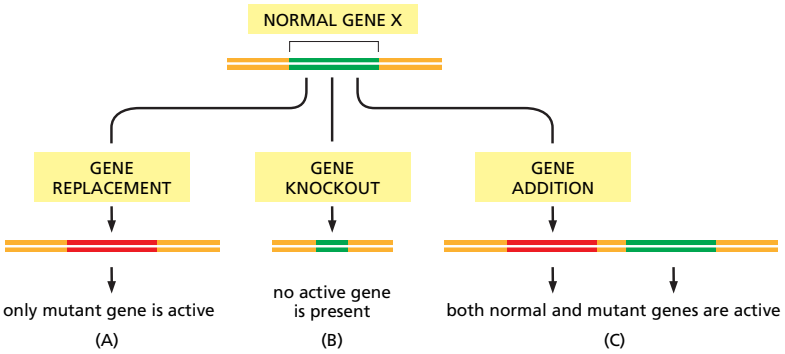



**Figure 8–64 Gene replacement, gene knockout, and gene addition.** A normal gene can be altered in several ways to produce a transgenic organism. (A) The normal gene *(green)* can be completely replaced by a mutant copy of the gene *(red)*. This provides information on the activity of the mutant gene without interference from the normal gene, and thus the effects of small and subtle mutations can be determined. (B) The normal gene can be inactivated completely, for example, by making a large deletion in it. (C) A mutant gene can simply be added to the genome. In some organisms this is the easiest type of genetic engineering to perform. This approach can provide useful information when the introduced mutant gene overrides the function of the normal gene, as with a dominant-negative mutation (see Figure 8–62).

Unlike higher eucaryotes (which are multicellular and diploid), bacteria, yeasts, and the cellular slime mold *Dictyostelium* generally exist as haploid single cells. In these organisms, an artificially introduced DNA molecule carrying a mutant gene can, with a relatively high frequency, replace the single copy of the normal gene by homologous recombination; it is therefore easy to produce cells in which the mutant gene has replaced the normal gene (**Figure 8–64**A). In this way, cells can be made in order to miss a particular protein or produce an altered form of it. The ability to perform direct gene replacements in lower eucaryotes, combined with the power of standard genetic analyses in these haploid organisms, explains in large part why studies in these types of cells have been so important for working out the details of those cell processes that are shared by all eucaryotes.

## Animals Can Be Genetically Altered

Gene additions and replacements are also possible, but more difficult to perform, in animals and plants. Animals and plants that have been genetically engineered by either gene insertion, gene deletion, or gene replacement are called **transgenic organisms**, and any foreign or modified genes that are added are called **transgenes**. We concentrate our discussion on transgenic mice, as enormous progress is being made in this area. If a DNA molecule carrying a mutated mouse gene is transferred into a mouse cell, it usually inserts into the chromosomes at random, but about once in a thousand times, it replaces one of the two copies of the normal gene by homologous recombination. By exploiting these rare "gene targeting" events, any specific gene can be altered or inactivated in a mouse cell by a direct gene replacement. In the special case in which both copies of the gene of interest is completely inactivated or deleted, the resulting animal is called a **"knockout" mouse**.

The technique works as follows. In the first step, a DNA fragment containing a desired mutant gene (or a DNA fragment designed to interrupt a target gene) is inserted into a vector and then introduced into cultured embryonic stem (ES) cells (see Figure 8–5), which are capable of producing cells of many different types. After a period of cell proliferation, the rare colonies of cells in which a homologous recombination event is likely to have caused a gene replacement to occur are isolated. The correct colonies among these are identified by PCR or by Southern blotting: they contain recombinant DNA sequences in which the inserted fragment has replaced all or part of one copy of the normal gene. In the second step, individual ES cells from the identified colony are taken up into a fine micropipette and injected into an early mouse embryo. The transfected ES cells collaborate with the cells of the host embryo to produce a normal-looking mouse; large parts of this chimeric animal, including—in favorable cases—cells of the germ line, often derive from the transfected ES cells (**Figure 8–65**).

The mice with the transgene in their germ line are then bred to produce both a male and a female animal, each heterozygous for the gene replacement (that is, they have one normal and one mutant copy of the gene). When these two

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 224 of 240

mice are mated, one-fourth of their progeny will be homozygous for the altered gene. Studies of these homozygotes allow the function of the altered gene—or the effects of eliminating the gene's activity—to be examined in the absence of the corresponding normal gene.

The ability to prepare transgenic mice lacking a known normal gene has been a major advance, and the technique is now being used to determine the functions of all mouse genes (**Figure 8–66**). A special technique is used to produce conditional mutants, in which a selected gene becomes disrupted in a specific tissue at a certain time in development. The strategy takes advantage of a site-specific recombination system to excise—and thus disable—the target gene in a particular place or at a particular time. The most common of these recombination systems, called **Cre/lox**, is widely used to engineer gene replacements in mice and in plants (see Figure 5–79). In this case, the target gene in ES cells is replaced by a fully functional version of the gene that is flanked by a pair of the

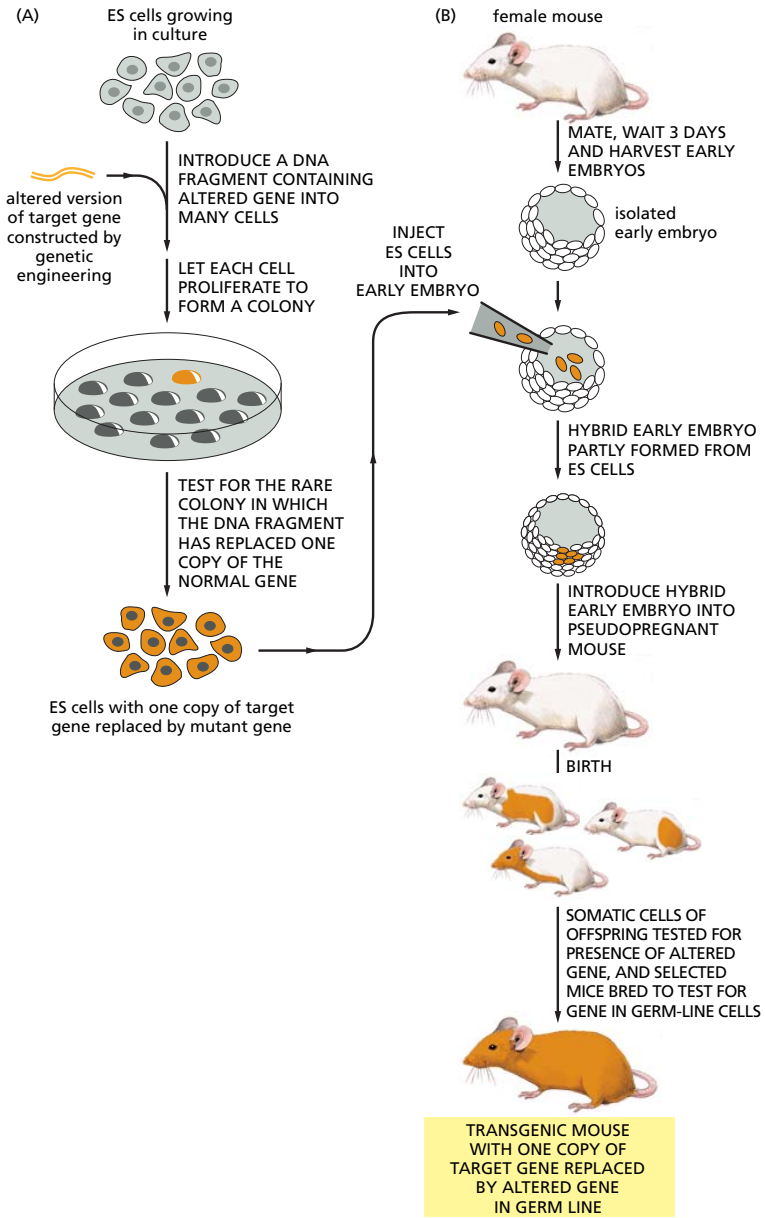
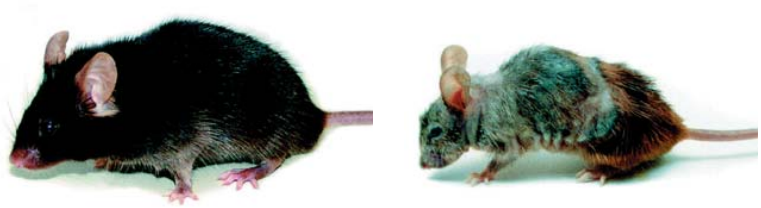



**Figure 8–65 Summary of the procedures used for making gene replacements in mice.** In the first step (A), an altered version of the gene is introduced into cultured ES (embryonic stem) cells. Only a few rare ES cells will have their corresponding normal genes replaced by the altered gene through a homologous recombination event. Although the procedure is often laborious, these rare cells can be identified and cultured to produce many descendants, each of which carries an altered gene in place of one of its two normal corresponding genes. In the next step of the procedure (B), these altered ES cells are injected into a very early mouse embryo; the cells are incorporated into the growing embryo, and a mouse produced by such an embryo will contain some somatic cells (indicated by *orange*) that carry the altered gene. Some of these mice will also contain germ-line cells that contain the altered gene; when bred with a normal mouse, some of the progeny of these mice will contain one copy of the altered gene in all of their cells. If two such mice are bred (not shown), some of the progeny will contain two altered genes (one on each chromosome) in all of their cells.

If the original gene alteration completely inactivates the function of the gene, these homozygous mice are known as knockout mice. When such mice are missing genes that function during development, they often die with specific defects long before they reach adulthood. These lethal defects are carefully analyzed to help determine the normal function of the missing gene.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 225 of 240



(A) (B)

**Figure 8–66 Transgenic mice engineered to express a mutant DNA helicase show premature aging.** The helicase, encoded by the *Xpd* gene, is involved in both transcription and DNA repair. Compared with a wild-type mouse of the same age (A), a transgenic mouse that expresses a defective version of *Xpd* (B) exhibits many of the symptoms of premature aging, including osteoporosis, emaciation, early graying, infertility, and reduced life-span. The mutation in *Xpd* used here impairs the activity of the helicase and mimics a mutation that in humans causes trichothiodystrophy, a disorder characterized by brittle hair, skeletal abnormalities, and a very reduced life expectancy. These results indicate that an accumulation of DNA damage can contribute to the aging process in both humans and mice. (From J. de Boer et al., *Science* 296:1276–1279, 2002. With permission from AAAS.)

short DNA sequences, called *lox sites,* that are recognized by the *Cre recombinase* protein. The transgenic mice that result are phenotypically normal. They are then mated with transgenic mice that express the Cre recombinase gene under the control of an inducible promoter. In the specific cells or tissues in which Cre is switched on, it catalyzes recombination between the lox sequences—excising a target gene and eliminating its activity. Similar recombination systems are used to generate conditional mutants in *Drosophila* (see Figure 22–49).

## Transgenic Plants Are Important for Both Cell Biology and Agriculture

A damaged plant can often repair itself by a process in which mature differentiated cells "dedifferentiate," proliferate, and then redifferentiate into other cell types. In some circumstances, the dedifferentiated cells can even form an apical meristem, which can then give rise to an entire new plant, including gametes. This remarkable developmental plasticity of plant cells can be exploited to generate transgenic plants from cells growing in culture.

When a piece of plant tissue is cultured in a sterile medium containing nutrients and appropriate growth regulators, many of the cells are stimulated to proliferate indefinitely in a disorganized manner, producing a mass of relatively undifferentiated cells called a *callus*. By carefully manipulating the nutrients and growth regulators, one can induce the formation of shoot and then root apical meristems within the callus, and, in many species, regenerate a whole new plant.

Callus cultures can also be mechanically dissociated into single cells, which will grow and divide as a suspension culture. In several plants—including tobacco, petunia, carrot, potato, and *Arabidopsis*—a single cell from such a suspension culture can be grown into a small clump (a clone) from which a whole plant can be regenerated. Such a cell, which has the ability to give rise to all parts of the organism, is considered **totipotent**. Just as mutant mice can be derived by the genetic manipulation of ES cells in culture, so transgenic plants can be created from single totipotent plant cells that have been transfected with DNA in culture (**Figure 8–67**).

The ability to produce transgenic plants has greatly accelerated progress in many areas of plant cell biology. It has had an important role, for example, in isolating receptors for growth regulators and in analyzing the mechanisms of morphogenesis and of gene expression in plants. It has also opened up many new possibilities in agriculture that could benefit both farmer and consumer. It has made it possible, for example, to modify the lipid, starch, and protein stored in seeds, to impart pest and virus resistance to plants, and to create modified plants that tolerate extreme habitats such as salt marshes or water-stressed soil.

Many of the major advances in understanding animal development have come from studies on the fruit fly *Drosophila* and the nematode worm *C. elegans,* which are amenable to classical genetic analysis, as well as to experimental manipulation. Progress in plant developmental biology has, in the past, been relatively slow by comparison. Many of the plants that have proved most amenable to genetic analysis—such as maize and tomato—have long life cycles and very large genomes, making both classical and molecular genetic analysis very time-consuming. Increasing attention is consequently being paid to a fast-growing small weed, the common wall cress *(Arabidopsis thaliana),* which has

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 226 of 240

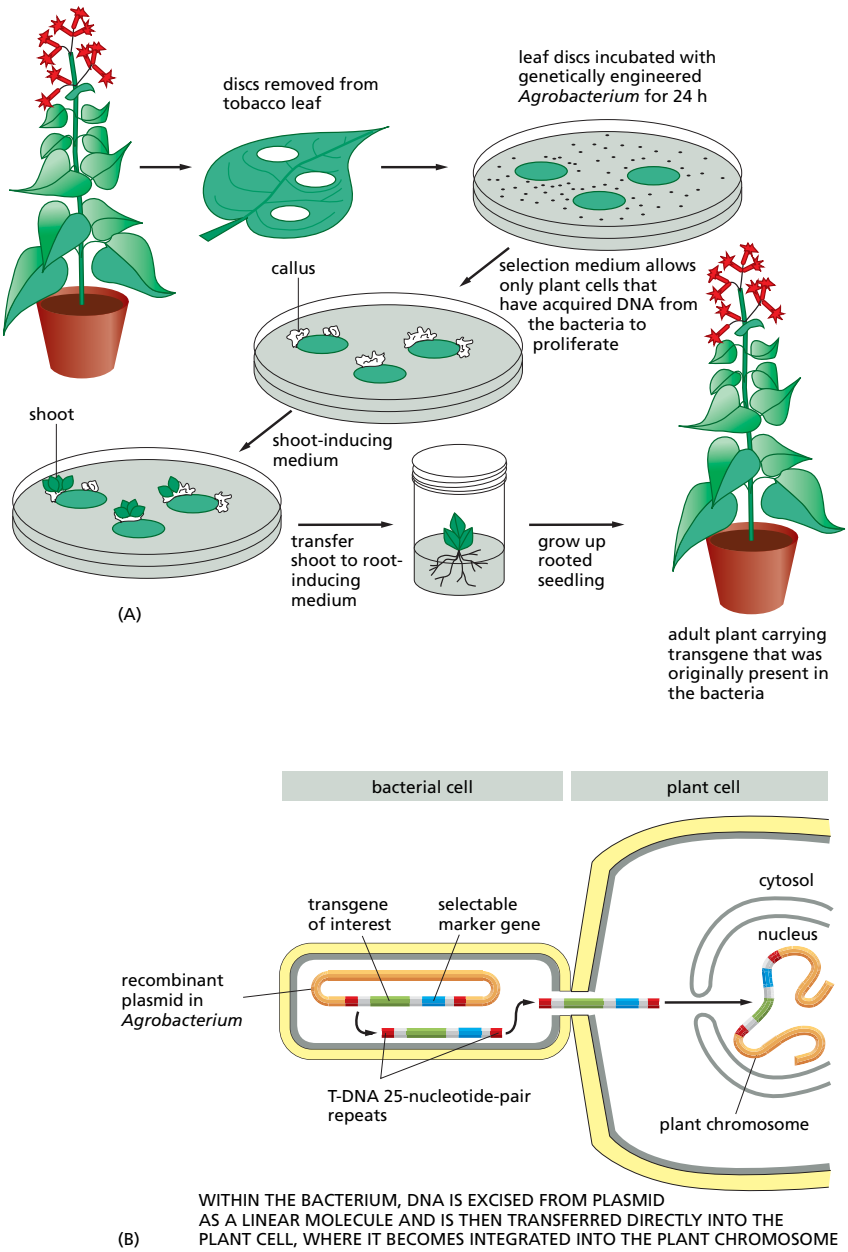



**Figure 8–67 A procedure used to make a transgenic plant.** (A) Outline of the process. A disc is cut out of a leaf and incubated in culture with *Agrobacterium* cells that carry a recombinant plasmid that contains both a selectable marker gene and a desired transgene. The wounded cells at the edge of the disc release substances that attract the *Agrobacterium* cells and cause them to inject DNA into these cells. Only those plant cells that take up the appropriate DNA and express the selectable marker gene survive to proliferate and form a callus. The manipulation of growth regulators and nutrients supplied to the callus induces it to form shoots, which subsequently root and grow into adult plants carrying the transgene. (B) The preparation of the recombinant plasmid and its transfer to plant cells. An *Agrobacterium* plasmid that normally carries the T-DNA sequence is modified by substituting a selectable marker gene (such as the kanamycin-resistance gene) and a desired transgene between the 25-nucleotide-pair T-DNA repeats. When the *Agrobacterium* recognizes a plant cell, it efficiently passes a DNA strand that carries these sequences into the plant cell, using the special machinery that normally transfers the plasmid's T-DNA sequence.

several major advantages as a "model plant" (see Figures 1–46 and 22–112). The relatively small *Arabidopsis* genome was the first plant genome to be completely sequenced, and the pace of research on this organism now rivals that of the model animals.

## Large Collections of Tagged Knockouts Provide a Tool for Examining the Function of Every Gene in an Organism

Extensive collaborative efforts are underway to assemble comprehensive libraries of mutations in a variety of model organisms, including *S. cerevisiae, C. elegans, Drosophila, Arabidopsis,* and the mouse. The ultimate aim in each case is to produce a collection of mutant strains in which every gene in the organism has been systematically deleted or altered in such a way that it can be conditionally disrupted. Collections of this type will provide an invaluable resource for investi-

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 227 of 240

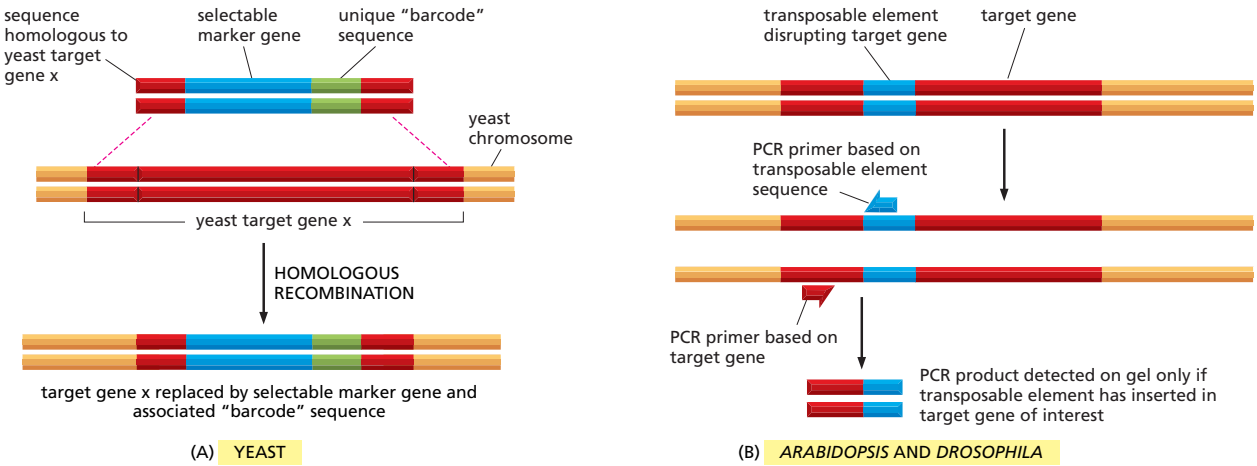



**Figure 8–68 Making collections of mutant organisms.** (A) A deletion cassette for use in yeast contains DNA sequences *(red)* homologous to each end of a target gene x, a selectable marker gene *(blue),* and a unique "barcode" sequence approximately 20 nucleotide pairs in length *(green)*. This DNA is introduced into yeast cells, where it readily replaces the target gene by homologous recombination. By using a collection of such cassettes, each specific for one gene, a library of yeast mutants can be constructed containing a mutant for every gene. (B) A similar approach can be taken to prepare tagged knockout mutants in *Arabidopsis* and *Drosophila*. In this case, mutations are generated by the accidental insertion of a transposable element into a target gene. The total DNA from the resulting organism can be collected and quickly screened for disruption of a gene of interest by using PCR primers that bind to the transposable element and to the target gene. A PCR product is detected on the gel only if the transposable element has inserted into the target gene (see Figure 8–45).

gating gene function on a genomic scale. In some cases, each of the individual mutations within the collection will express a distinct molecular tag—in the form of a unique DNA sequence—designed to make identification of the altered gene rapid and routine.

In *S. cerevisiae,* the task of generating a complete set of 6000 mutants, each missing only one gene, is made simpler by yeast's propensity for homologous recombination. For each gene, a "deletion cassette" is prepared. The cassette consists of a special DNA molecule that contains 50 nucleotides identical in sequence to each end of the targeted gene, surrounding a selectable marker. In addition, a special "barcode" sequence tag is embedded in this DNA molecule to facilitate the later rapid identification of each resulting mutant strain (**Figure 8–68**). A large mixture of such gene knockout mutants can then be grown under various selective test conditions—such as nutritional deprivation, a temperature shift, or the presence of various drugs—and the cells that survive can be rapidly identified by their unique sequence tags. By assessing how well each mutant in the mixture fares, one can begin to assess which genes are essential, useful, or irrelevant for growth under the various conditions.

The challenge in deriving information from the study of such yeast mutants lies in deducing a gene's activity or biological role based on a mutant phenotype. Some defects—an inability to live without histidine, for example—point directly to the function of the wild-type gene. Other connections may not be so obvious. What might a sudden sensitivity to cold indicate about the role of a particular gene in the yeast cell? Such problems are even greater in organisms that are more complex than yeast. The loss of function of a single gene in the mouse, for example, may affect many different tissue types at different stages of development—whereas the loss of other genes may have no obvious effect. Adequately characterizing mutant phenotypes in mice often requires a thorough examination, along with extensive knowledge of mouse anatomy, histology, pathology, physiology, and complex behavior.

The insights generated by examination of mutant libraries, however, will be great. For example, studies of an extensive collection of mutants in *Mycoplasma genitalium*—the organism with the smallest known genome—have identified the minimum complement of genes essential for cellular life. Analysis of the mutant pool suggests that growth under laboratory conditions requires about three-quarters of the 480 protein-coding genes in *M. genitalium*. Approximately

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 228 of 240

100 of these essential genes are of unknown function, which suggests that a surprising number of the basic molecular mechanisms that underlie life have yet to be discovered.

## RNA Interference Is a Simple and Rapid Way to Test Gene Function

Although knocking out a gene in an organism and studying the consequences is perhaps the most powerful approach for understanding the functions of the gene, a much easier way to inactivate genes has been recently discovered. Called **RNA interference** (**RNAi**, for short), this method exploits a natural mechanism used in many plants, animals, fungi, and protozoa to protect themselves against certain viruses and transposable elements (see Figure 7–115). The technique introduces into a cell or organism a double-stranded RNA molecule whose nucleotide sequence matches that of part of the gene to be inactivated. After the RNA is processed, it hybridizes with the mRNA produced by the target gene and directs its degradation. The cell subsequently uses small fragments of this degraded RNA to produce more double-stranded RNA, which directs the continued elimination of the target mRNA. Because these short RNA fragments can be passed on to progeny cells, RNAi can cause heritable changes in gene expression. But, as we saw in Chapter 7, there is a second mechanism through which RNAi can stably inactivate genes. RNA fragments produced by degradation in the cytosol can enter the nucleus and interact with the target gene itself, directing its packaging into a transcriptionally repressed form of chromatin. This dual mode of controlling gene expression makes RNAi an especially effective tool for shutting down genes, one at a time.

RNAi is frequently used to inactivate genes in *Drosophila* and mammalian cell culture lines. Indeed, sets of 15,000 *Drosophila* RNAi molecules (one for every gene) allow researchers, in several months, to test the role of every fly gene in any process that can be monitored using cultured cells. Soon, it will be possible to carry out the same type of analysis with the 25,000 mouse and human genes. RNAi has also been widely used to study gene function in the nematode, *C. elegans*. When working with worms, introducing the double-stranded RNA is quite simple: the RNA can be injected directly into the intestine of the animal, or the worm can be fed with *E. coli* engineered to produce the RNA (**Figure 8–69**). The RNA is distributed throughout the body of the worm, where it inhibits expression of the target gene in different tissue types. Because the entire genome of *C. elegans* has been sequenced, RNAi is being used to help in assigning functions to the entire complement of worm genes.

More recently, a related technique has also been widely applied to mice. In this case, the RNAi is not injected or fed to the mouse; rather, recombinant DNA techniques are used to make transgenic animals that express the RNAi under the control of an inducible promoter. Often this is a specially designed RNA that can fold back on itself and, through base pairing, produce a double-stranded region

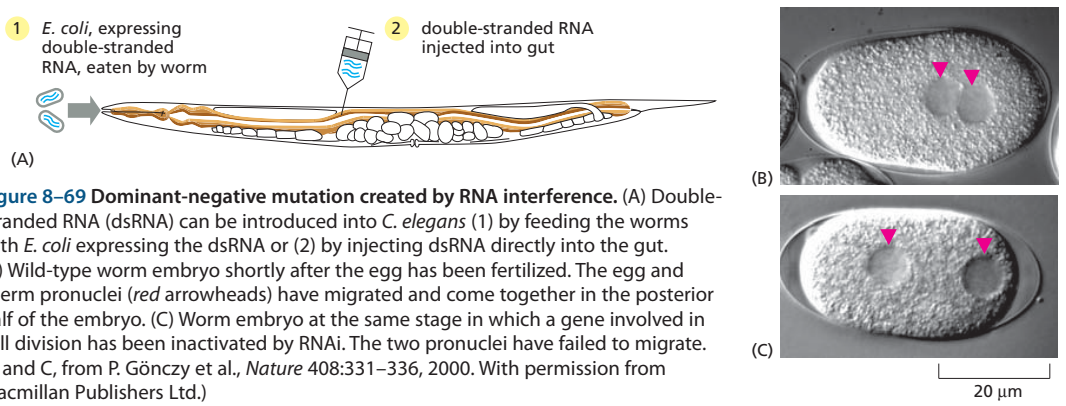



**Figure 8–69 Dominant-negative mutation created by RNA interference.** (A) Double-stranded RNA (dsRNA) can be introduced into *C. elegans* (1) by feeding the worms with *E. coli* expressing the dsRNA or (2) by injecting dsRNA directly into the gut. (B) Wild-type worm embryo shortly after the egg has been fertilized. The egg and sperm pronuclei (*red* arrowheads) have migrated and come together in the posterior half of the embryo. (C) Worm embryo at the same stage in which a gene involved in cell division has been inactivated by RNAi. The two pronuclei have failed to migrate. (B and C, from P. Gönczy et al., *Nature* 408:331–336, 2000. With permission from Macmillan Publishers Ltd.)

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 229 of 240

that is recognized by the RNAi machinery. The process inactivates only the genes that exactly match the RNAi sequence. Depending on the inducible promoter used, the RNAi can be produced only in a specified tissue or only at a particular time in development, allowing the functions of the target genes to be analyzed in elaborate detail.

RNAi has made reverse genetics simple and efficient in many organisms, but it has several potential limitations compared with true genetic knockouts. For unknown reasons, RNAi does not efficiently inactivate all genes. Moreover, within whole organisms, certain tissues may be resistant to the action of RNAi (for example, neurons in nematodes). Another problem arises because many organisms contain large gene families, the members of which exhibit sequence similarity. RNAi therefore sometimes produces "off-target" effects, inactivating related genes in addition to the targeted gene. One strategy to avoid such problems is to use multiple small RNA molecules matched to different regions of the same gene. Ultimately, the results of any RNAi experiment must be viewed as a strong clue to, but not necessarily a proof of, normal gene function.

## Reporter Genes and *In Situ* Hybridization Reveal When and Where a Gene Is Expressed

Important insights into gene function can often be obtained by examining when and where a gene is expressed in the cell or in the whole organism. Determining the pattern and timing of gene expression can be accomplished by replacing the coding portion of the gene under study with a reporter gene. In most cases, the expression of the reporter gene is then monitored by tracking the fluorescence or enzymatic activity of its protein product (see Figures 9–26 and 9–27).

As discussed in detail in Chapter 7, regulatory DNA sequences, located upstream or downstream of the coding region, control gene expression. These regulatory sequences, which determine exactly when and where the gene is expressed, can be easily studied by placing a reporter gene under their control and introducing these recombinant DNA molecules into cells (**Figure 8–70**).

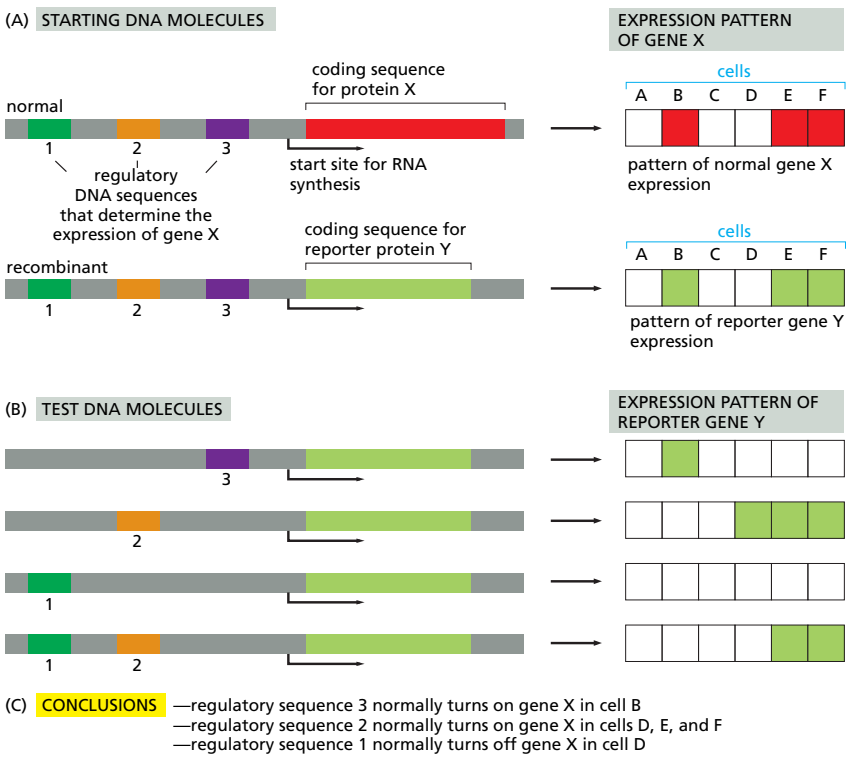



**Figure 8–70 Using a reporter protein to determine the pattern of a gene's expression.** (A) In this example, the coding sequence for protein X is replaced by the coding sequence for reporter protein Y. The expression pattern for X and Y are the same. (B) Various fragments of DNA containing candidate regulatory sequences are added in combinations to produce test DNA molecules encoding reporter gene Y. These recombinant DNA molecules are then tested for expression after their transfection into a variety of different types of mammalian cells. The results are summarized in (C).

For experiments in eucaryotic cells, two commonly used reporter proteins are the enzyme β-galactosidase *(β-gal)* (see Figure 7–55B) and green fluorescent protein or GFP (see Figure 9–26). Figure 7–55B shows an example in which the *β-gal* gene is used to monitor the activity of the *Eve* gene regulatory sequence in a *Drosophila* embryo.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 230 of 240

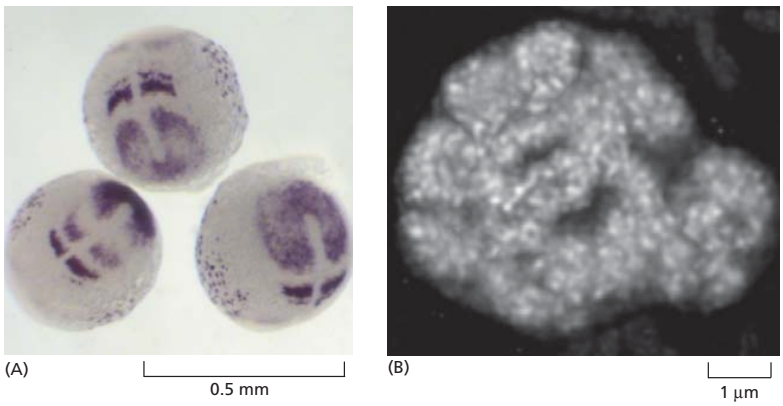



**Figure 8–71 *In situ* hybridization for RNA localization.** (A) Expression pattern of *DeltaC* mRNA in the early zebrafish embryo. This gene codes for a ligand in the Notch signaling pathway (discussed in Chapter 15), and the pattern shown here reflects its role in the development of somites—the future segments of the vertebrate trunk and tail. (B) High-resolution RNA *in situ* localization reveals the sites within the nucleolus of a pea cell where ribosomal RNA is synthesized. The sausage-like structures, 0.5–1 μm in diameter, correspond to the loops of chromosomal DNA that contain the genes encoding rRNA. Each small white spot represents transcription of a single rRNA gene. (A, courtesy of Yun-Jin Jiang; B, courtesy of Peter Shaw.)

It is also possible to directly observe the time and place that the mRNA product of a gene is expressed. Although this strategy often provides the same general information as the reporter gene approaches discussed above, there are instances where it provides additional information; for example, when the gene is transcribed but the mRNA is not immediately translated, or when the gene's final product is RNA rather than protein. This procedure, called ***in situ* hybridization**, relies on the principles of nucleic acid hybridization described earlier. Typically, tissues are gently fixed so that their RNA is retained in an exposed form that can hybridize with a labeled complementary DNA or RNA probe. In this way, the patterns of differential gene expression can be observed in tissues, and the location of specific RNAs in cells can be determined (**Figure 8–71**). In the *Drosophila* embryo, for example, such patterns have provided new insights into the mechanisms that create distinctions between cells in different positions during development (described in Chapter 22).

Using similar approaches, it is also possible to visualize specific DNA sequences in cells. In this case, tissue, cell, or even chromosome preparations are briefly exposed to high pH to disrupt their nucleotide pairs, and nucleic acid probes are added, allowed to hybridize with the cells' DNA, and then visualized (see Figure 8–35).

## Expression of Individual Genes Can Be Measured Using Quantitative RT-PCR

Although reporter genes and *in situ* hybridization reveal patterns of gene expression, it is often desirable to quantitate gene expression by directly measuring mRNA levels in cells. Although Northern blots (see Figure 8–38) can be adapted to this purpose, a more accurate method is based on the principles of PCR (**Figure 8–72**). This method, called **quantitative RT-PCR** (reverse transcription-polymerase chain reaction), begins with the total population of mRNA molecules purified from a tissue or a cell culture. It is important that no DNA be present in the preparation; it must be purified away or enzymatically degraded. Two DNA primers that specifically match the gene of interest are added, along with reverse transcriptase, DNA polymerase, and the four deoxynucleoside triphosphates needed for DNA synthesis. The first round of synthesis is the reverse transcription of the mRNA into DNA using one of the primers. Next, a series of heating and cooling cycles allows the amplification of that DNA strand by conventional PCR (see Figure 8–45). The quantitative part of this method relies on a direct relationship between the rate at which the PCR product is generated and the original concentration of the mRNA species of interest. By adding chemical dyes to the PCR reaction that fluoresce only when bound to double-stranded DNA, a simple fluorescence measurement can be used to track the progress of the reaction and thereby accurately deduce the starting concentration of the mRNA that is amplified (see Figure 8–72). Although it seems complicated, this quantitative RT-PCR technique (sometimes called *real time PCR*) is relatively fast and simple

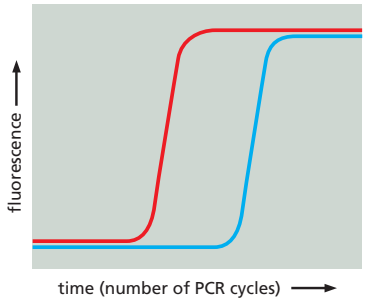



**Figure 8–72 RNA levels can be measured by quantitative RT-PCR.** The fluorescence measured is generated by a dye that fluoresces only when bound to the double-stranded DNA products of the RT-PCR reaction (see Figure 8–46B). The red sample has a higher concentration of the mRNA being measured than does the blue sample, since it requires fewer PCR cycles to reach the same half-maximal concentration of double-stranded DNA. Based on this difference, the relative amounts of the mRNA in the two samples can be precisely determined.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 231 of 240

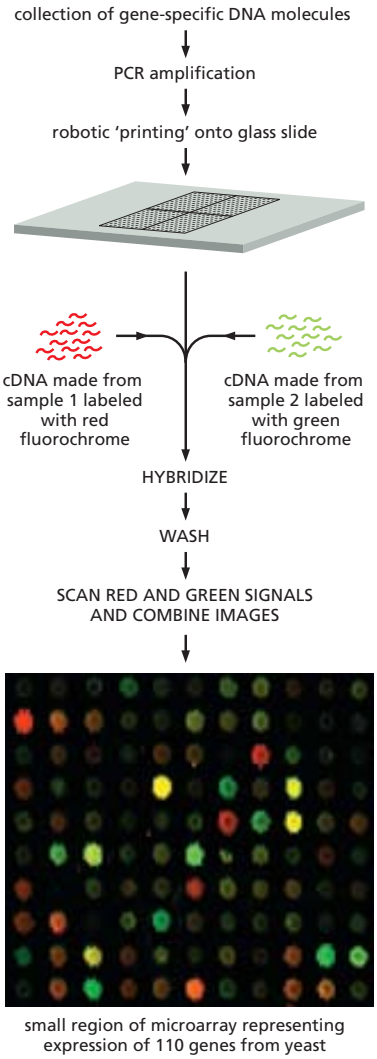



**Figure 8–73 Using DNA microarrays to monitor the expression of thousands of genes simultaneously.** To prepare the microarray, DNA fragments—each corresponding to a gene—are spotted onto a slide by a robot. Prepared arrays are also widely available commercially. In this example, mRNA is collected from two different cell samples for a direct comparison of their relative levels of gene expression; the two samples, for example, could be from cells treated with a hormone and untreated cells of the same type. These samples are converted to cDNA and labeled, one with a red fluorochrome, the other with a green fluorochrome. The labeled samples are mixed and then allowed to hybridize to the microarray. After incubation, the array is washed and the fluorescence scanned. In the portion of a microarray shown, which represents the expression of 110 yeast genes, *red* spots indicate that the gene in sample 1 is expressed at a higher level than the corresponding gene in sample 2; *green* spots indicate that expression of the gene is higher in sample 2 than in sample 1. *Yellow* spots reveal genes that are expressed at equal levels in both cell samples. Dark spots indicate little or no expression in either sample of the gene whose fragment is located at that position in the array. (Microarray courtesy of J.L. DeRisi et al., *Science* 278:680–686, 1997. With permission from AAAS.)

to perform in the laboratory; it has displaced Northern blotting as the method of choice for quantifying mRNA levels from any given gene.

## Microarrays Monitor the Expression of Thousands of Genes at Once

So far we have discussed techniques that can be used to monitor the expression of only a single gene (or relatively few genes) at a time. Developed in the 1990s, **DNA microarrays** have revolutionized the analysis of gene expression by monitoring the RNA products of thousands of genes at once. By examining the expression of so many genes simultaneously, we can now begin to identify and study the gene expression patterns that underlie cell physiology: we can see which genes are switched on (or off) as cells grow, divide, differentiate, or respond to hormones or to toxins.

DNA microarrays are little more than glass microscope slides studded with a large number of DNA fragments, each containing a nucleotide sequence that serves as a probe for a specific gene. The most dense arrays may contain tens of thousands of these fragments in an area smaller than a postage stamp, allowing thousands of hybridization reactions to be performed in parallel (**Figure 8–73**). Some microarrays are prepared from large DNA fragments that have been generated by PCR and then spotted onto the slides by a robot. Others contain short oligonucleotides that are synthesized on the surface of the glass wafer with techniques similar to those that are used to etch circuits onto computer chips. In either case, the exact sequence—and position—of every probe on the chip is known. Thus, any nucleotide fragment that hybridizes to a probe on the array can be identified as the product of a specific gene simply by detecting the position at which it is bound.

To use a DNA microarray to monitor gene expression, mRNA from the cells being studied is first extracted and converted to cDNA (see Figure 8–43). The cDNA is then labeled with a fluorescent probe. The microarray is incubated with this labeled cDNA sample and hybridization is allowed to occur (see Figure 8–73). The array is then washed to remove cDNA that is not tightly bound, and the positions in the microarray to which labeled DNA fragments have bound are identified by an automated scanning-laser microscope. The array positions are then matched to the particular gene whose sample of DNA was spotted in this location.

Typically the fluorescent DNA from the experimental samples (labeled, for example, with a red fluorescent dye) are mixed with a reference sample of cDNA fragments labeled with a differently colored fluorescent dye (green, for example). Thus, if the amount of RNA expressed from a particular gene in the cells of interest is increased relative to that of the reference sample, the resulting spot is red. Conversely, if the gene's expression is decreased relative to the reference sample, the spot is green. If there is no change compared to the reference sample, the spot

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 232 of 240

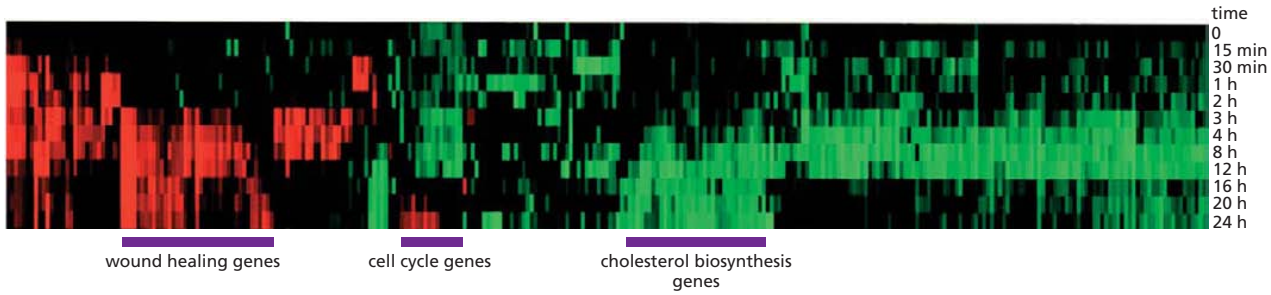



**Figure 8–74 Using cluster analysis to identify sets of genes that are coordinately regulated.** Genes that belong to the same cluster may be involved in common pathways or processes. To perform a cluster analysis, microarray data are obtained from cell samples exposed to a variety of different conditions, and genes that show coordinate changes in their expression pattern are grouped together. In this experiment, human fibroblasts were deprived of serum for 48 hours; serum was then added back to the cultures at time 0 and the cells were harvested for microarray analysis at different time points. Of the 8600 genes analyzed on the DNA microarray, just over 300 showed threefold or greater variation in their expression patterns in response to serum re-introduction. Here, *red* indicates an increase in expression; *green* is a decrease in expression. On the basis of the results of many microarray experiments, the 8600 genes have been grouped in clusters based on similar patterns of expression. The results of this analysis show that genes involved in wound healing are turned on in response to serum, while genes involved in regulating cell cycle progression and cholesterol biosynthesis are shut down. (From M.B. Eisen et al., *Proc. Natl Acad. Sci. U.S.A.* 95:14863–14868, 1998. With permission from National Academy of Sciences.)

is yellow. Using such an internal reference, gene expression profiles can be tabulated with great precision.

So far, DNA microarrays have been used to examine everything from the changes in gene expression that make strawberries ripen to the gene expression "signatures" of different types of human cancer cells (see Figure 7–3); or from changes that occur as cells progress through the cell cycle to those made in response to sudden shifts in temperature. Indeed, because microarrays allow the simultaneous monitoring of large numbers of genes, they can detect subtle changes in a cell, changes that might not be manifested in its outward appearance or behavior.

Comprehensive studies of gene expression also provide an additional layer of information that is useful for predicting gene function. Earlier, we discussed how identifying a protein's interaction partners can yield clues about that protein's function. A similar principle holds true for genes: information about a gene's function can be deduced by identifying genes that share its expression pattern. Using a technique called *cluster analysis*, one can identify sets of genes that are coordinately regulated. Genes that are turned on or turned off together under different circumstances are likely to work in concert in the cell: they may encode proteins that are part of the same multiprotein machine, or proteins that are involved in a complex coordinated activity, such as DNA replication or RNA splicing. Characterizing a gene whose function is unknown by grouping it with known genes that share its transcriptional behavior is sometimes called "guilt by association." Cluster analyses have been used to analyze the gene expression profiles that underlie many interesting biological processes, including wound healing in humans (**Figure 8–74**).

In addition to monitoring the level of mRNA corresponding to every gene in a genome, DNA microarrays have many other uses. For example, they can be used to monitor the progression of DNA replication in a cell (see Figure 5–32) and, when combined with immunoprecipitation, can pinpoint every position in the genome occupied by a given gene regulatory protein (see Figure 7–32). Microarrays can also be used to quickly identify disease-causing microbes by hybridizing DNA from infected tissues to an array containing genomic DNA sequences from large collections of pathogens.

## Single-Cell Gene Expression Analysis Reveals Biological "Noise"

The methods for monitoring mRNAs just described give average expression levels for each mRNA across a large population of cells. By using a fluorescent

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 233 of 240

reporter protein whose expression is under the control of a promoter of interest, it is also possible to accurately measure expression levels in individual cells. These new approaches have revealed a startling amount of variability, often called *biological noise,* between the individual cells in a homogeneous population of cells. These studies have also revealed the presence of distinct subpopulations of cells whose existence would be masked if only the average across a whole population were considered. For example, a bimodal distribution of expression levels would indicate that the cells can exist in two distinct states (**Figure 8–75**), with the average expression level of the population being somewhere between them. The behavior of individual cells has important implications for understanding biology, for example, by revealing that some cells constantly and rapidly switch back and forth between two states.

Currently, there are two approaches for monitoring gene expression in individual cells. In the imaging approach, live cells are mounted on a slide and viewed through a fluorescence microscope. This method has the advantage that a given cell can be followed over time, allowing temporal changes in expression to be measured. The second approach, flow cytometry, works by streaming a dilute suspension of cells past an illuminator and measuring the fluorescence of individual cells as they flow past the detector (see Figure 8–2). Although it has the advantage that the expression levels of very large numbers of cells can be measured with precision, flow cytometry does not allow a given cell to be tracked over time; hence, it is complementary to the imaging methods.

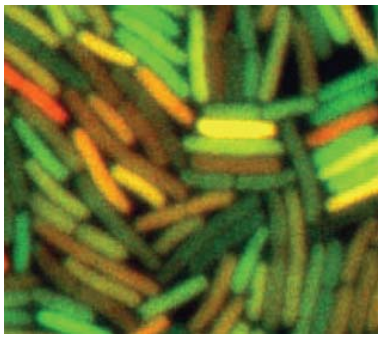


**Figure 8–75 Different levels of gene expression in individual cells within a population of *E. coli* bacteria.** For this experiment, two different reporter proteins (one fluorescing *green,* the other *red*) controlled by a copy of the same promoter, have been introduced into all of the bacteria. When illuminated, some cells express only one gene copy, and so appear either *red* or *green,* while others express both gene copies, and so appear *yellow.* This experiment also reveals variable levels of fluorescence, indicating variable levels of gene expression within an apparently uniform population of cells. (From M.B. Elowitz, A.J. Levine, E.O. Siggia and P.S. Swain, *Science* 297:1183–1186, 2002. With permission from AAAS.)

## Summary

*Genetics and genetic engineering provide powerful tools for the study of gene function in both cells and organisms. In the classical genetic approach, random mutagenesis is coupled with screening to identify mutants that are deficient in a particular biological process. These mutants are then used to locate and study the genes responsible for that process.*

*Gene function can also be ascertained by reverse genetic techniques. DNA engineering methods can be used to alter genes and to re-insert them into a cell's chromosomes so that they become a permanent part of the genome. If the cell used for this gene transfer is a fertilized egg (for an animal) or a totipotent plant cell in culture, transgenic organisms can be produced that express the mutant gene and pass it on to their progeny. Especially important for cell biology is the ability to alter cells and organisms in highly specific ways—allowing one to discern the effect on the cell or the organism of a designed change in a single protein or RNA molecule.*

*Many of these methods are being expanded to investigate gene function on a genome-wide scale. The generation of mutant libraries in which every gene in an organism has been systematically deleted or disrupted provides invaluable tools for exploring the role of each gene in the elaborate molecular collaboration that gives rise to life. Technologies such as DNA microarrays can monitor the expression of thousands of genes simultaneously, providing detailed, comprehensive snapshots of the dynamic patterns of gene expression that underlie complex cell processes.*

# PROBLEMS

### Which statements are true? Explain why or why not.

**8–1** Because a monoclonal antibody recognizes a specific antigenic site (epitope), it binds only to the specific protein against which it was made.

**8–2** Given the inexorable progress of technology, it seems inevitable that the sensitivity of detection of molecules will ultimately be pushed beyond the yoctomole level ($10^{-24}$ mole).

**8–3** Surface plasmon resonance (SPR) measures association ($k_{on}$) and dissociation ($k_{off}$) rates between molecules in real time, using small amounts of unlabeled molecules, but it does not give the information needed to determine the binding constant ($K$).

**8–4** If each cycle of PCR doubles the amount of DNA synthesized in the previous cycle, then 10 cycles will give a $10^3$-fold amplification, 20 cycles will give a $10^6$-fold amplification, and 30 cycles will give a $10^9$-fold amplification.

### Discuss the following problems.

**8–5** A common step in the isolation of cells from a sample of animal tissue is to treat it with trypsin, collagenase, and EDTA. Why is such a treatment necessary, and what

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 234 of 240

does each component accomplish? And why does this treatment not kill the cells?

**8–6** Do you suppose it would be possible to raise an antibody against another antibody? Explain your answer.

**8–7** Distinguish between velocity sedimentation and equilibrium sedimentation. For what general purpose is each technique used? Which do you suppose might be best suited for separating two proteins of different size?

**8–8** Tropomyosin, at 93 kd, sediments at 2.6 S, whereas the 65-kd protein, hemoglobin, sediments at 4.3 S. (The sedimentation coefficient S is a linear measure of the rate of sedimentation: both increase or decrease in parallel.) These two proteins are shown as α-carbon backbone models in **Figure Q8–1**. How is it that the bigger protein sediments more slowly than the smaller one? Can you think of an analogy from everyday experience that might help you with this problem?

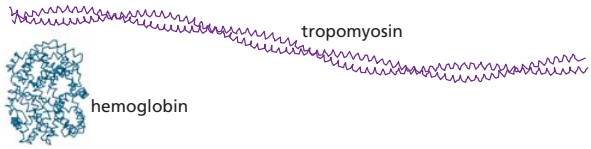



**Figure Q8–1** Backbone models of tropomyosin and hemoglobin (Problem 8–8).

**8–9** In the classic paper that demonstrated the semi-conservative replication of DNA, Meselson and Stahl began by showing that DNA itself will form a band when subjected to equilibrium sedimentation. They mixed randomly fragmented *E. coli* DNA with a solution of CsCl so that the final solution had a density of 1.71 g/mL. As shown in **Figure Q8–2**, with increasing length of centrifugation at 70,000 times gravity, the DNA, which was initially dispersed throughout the centrifuge tube, became concentrated over time into a discrete band in the middle.

**A.** Describe what is happening with time and explain why the DNA forms a discrete band.

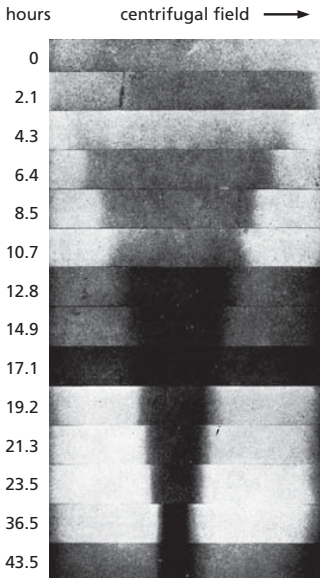



**Figure Q8–2** Ultraviolet absorption photographs showing successive stages in the banding of *E. coli* DNA (Problem 8–9). DNA, which absorbs UV light, shows up as *dark regions* in the photographs. The bottom of the centrifuge tube is on the *right*. (From M. Meselson and F.W. Stahl, *Proc. Natl Acad. Sci. U.S.A.* 44:671–682, 1958. With permission from National Academy of Sciences.)

**B.** What is the buoyant density of the DNA? (The density of the solution at which DNA "floats" at equilibrium defines the "buoyant density" of the DNA.)

**C.** Even if the DNA were centrifuged for twice as long—or even longer—the width of the band remains about what is shown at the bottom of Figure Q8–2. Why does the band not become even more compressed? Suggest some possible reasons to explain the thickness of the DNA band at equilibrium.

**8–10** Hybridoma technology allows one to generate monoclonal antibodies to virtually any protein. Why is it then that tagging proteins with epitopes is such a commonly used technique, especially since an epitope tag has the potential to interfere with the function of the protein?

**8–11** How many copies of a protein need to be present in a cell in order for it to be visible as a band on a gel? Assume that you can load 100 μg of cell extract onto a gel and that you can detect 10 ng in a single band by silver staining. The concentration of protein in cells is about 200 mg/mL, and a typical mammalian cell has a volume of about 1000 $\mu m^3$ and a typical bacterium a volume of about 1 $\mu m^3$. Given these parameters, calculate the number of copies of a 120-kd protein that would need to be present in a mammalian cell and in a bacterium in order to give a detectable band on a gel. You might try an order-of-magnitude guess before you make the calculations.

**8–12** You want to amplify the DNA between the two stretches of sequence shown in **Figure Q8–3**. Of the listed primers choose the pair that will allow you to amplify the DNA by PCR.

DNA to be amplified

```
5′-GACCTGTGGAAGC ———————— CATACGGGATTGA-3′
3′-CTGGACACCTTCG ———————— GTATGCCCTAACT-5′
```

primers

| | |
|---|---|
| (1) 5′-GACCTGTCCAAGC-3′ | (5) 5′-CATACGGGATTGA-3′ |
| (2) 5′-CTGGACACCTTCG-3′ | (6) 5′-GTATGCCCTAACT-3′ |
| (3) 5′-CGAAGGTGTCCAG-3′ | (7) 5′-TGTTAGGGCATAC-3′ |
| (4) 5′-GCTTCCACAGGTC-3′ | (8) 5′-TCAATCCCGTATG-3′ |

**Figure Q8–3** DNA to be amplified and potential PCR primers (Problem 8–12).

**8–13** In the very first round of PCR using genomic DNA, the DNA primers prime synthesis that terminates only when the cycle ends (or when a random end of DNA is encountered). Yet, by the end of 20 to 30 cycles—a typical amplification—the only visible product is defined precisely by the ends of the DNA primers. In what cycle is a double-stranded fragment of the correct size first generated?

**8–14** Explain the difference between a gain-of-function mutation and a dominant-negative mutation. Why are both these types of mutation usually dominant?

**8–15** Discuss the following statement: "We would have no idea today of the importance of insulin as a regulatory hormone if its absence were not associated with the devastating human disease diabetes. It is the dramatic consequences of its absence that focused early efforts on the identification of insulin and the study of its normal role in physiology."

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 235 of 240

# REFERENCES


### General

Ausubel FM, Brent R, Kingston RE et al (eds) (2002) Short Protocols in Molecular Biology, 5th ed. New York: Wiley.

Brown TA (2002) Genomes 2, 2nd ed. New York: Wiley-Liss.

Spector DL, Goldman RD & Leinwand LA (eds) (1998) *Cells:* A Laboratory Manual. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Watson JD, Caudy AA, Myers RM & Witkowski JA (2007) Recombinant DNA: Genes and Genomes—A Short Course, 3rd ed. New York: WH Freeman.

### Isolating Cells and Growing Them in Culture

Emmert-Buck MR, Bonner RF, Smith PD et al (1996) Laser capture microdissection. *Science* 274:998–1001.

Ham RG (1965) Clonal growth of mammalian cells in a chemically defined, synthetic medium. *Proc Natl Acad Sci USA* 53:288–293.

Harlow E & Lane D (1999) *Using Antibodies: A Laboratory Manual*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Herzenberg LA, Sweet RG & Herzenberg LA (1976) Fluorescence-activated cell sorting. *Sci Am* 234:108–116.

Levi-Montalcini R (1987) The nerve growth factor thirty-five years later. *Science* 237:1154–1162.

Lerou PH & Daley GQ (2005) Therapeutic potential of embryonic stem cells. *Blood Rev* 19:321–31.

Milstein C (1980) Monoclonal antibodies. *Sci Am* 243:66–74.

### Purifying Proteins

de Duve C & Beaufay H (1981) A short history of tissue fractionation. *J Cell Biol* 91:293s–299s.

Krogan NJ, Cagney G, Yu H et al (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* 440:637–43.

Laemmli UK (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* 227:680–685.

Nirenberg MW & Matthaei JH (1961) The dependence of cell-free protein synthesis in *E. coli* on naturally occurring or synthetic polyribonucleotides. *Proc Natl Acad Sci. USA* 47:1588–1602.

O'Farrell PH (1975) High-resolution two-dimensional electrophoresis of proteins. *J Biol Chem* 250:4007–4021.

Palade G (1975) Intracellular aspects of the process of protein synthesis. *Science* 189:347–358.

Scopes RK & Cantor CR (1994) Protein Purification: Principles and Practice, 3rd ed. New York: Springer-Verlag.

### Analyzing Proteins

Branden C & Tooze J (1999) Introduction to Protein Structure, 2nd ed. New York: Garland Science.

Fields S & Song O (1989) A novel genetic system to detect protein–protein interactions. *Nature* 340:245–246.

Giepmans BN, Adams SR et al (2006) The fluorescent toolbox for assessing protein location and function. *Science* 312:217–24.

Kendrew JC (1961) The three-dimensional structure of a protein molecule. *Sci Am* 205:96–111.

Knight ZA & Shokat KM (2007) Chemical genetics: Where genetics and pharmacology meet. *Cell* 128:425–30.

Rigaut G, Shevchenko A, Rutz B et al (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnol* 17:1030–1032.

Washburn MP, Wolters D and Yates JR (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnol* 19:242–7.

Wuthrich K (1989) Protein structure determination in solution by nuclear magnetic resonance spectroscopy. *Science* 243:45–50.

### Isolating, Cloning, and Sequencing DNA

Adams MD, Celniker SE, Holt RA et al (2000) The genome sequence of *Drosophila melanogaster*. *Science* 287:2185–2195.

Alwine JC, Kemp DJ & Stark GR (1977) Method for detection of specific RNAs in agarose gels by transfer to diabenzyloxymethyl-paper and hybridization with DNA probes. *Proc Natl Acad Sci USA* 74:5350–5354.

Blattner FR, Plunkett G, Bloch CA et al (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* 277:1453–1474.

Cohen S, Chang A, Boyer H & Helling R (1973) Construction of biologically functional bacterial plasmids *in vitro*. *Proc Natl Acad Sci USA* 70:3240–3244.

International Human Genome Sequencing Consortium (2000) Initial sequencing and analysis of the human genome. *Nature* 409:860–921.

International Human Genome Sequencing Consortium (2006) The DNA sequence, annotation and analysis of human chromosome 3. *Nature* 440:1194–1198.

Jackson D, Symons R & Berg P (1972) Biochemical method for inserting new genetic information into DNA of simian virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of *Escherichia coli*. *Proc Natl Acad Sci USA* 69:2904–2909.

Maniatis T et al (1978) The isolation of structural genes from libraries of eukaryotic DNA. *Cell* 15:687–701.

Mullis KB (1990). The unusual origin of the polymerase chain reaction. *Sci Am* 262:56–61.

Nathans D & Smith HO (1975) Restriction endonucleases in the analysis and restructuring of dna molecules. *Annu Rev Biochem* 44:273–93.

Saiki RK, Gelfand DH, Stoffel S et al (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* 239:487–491.

Sambrook J, Russell D (2001) Molecular Cloning: A Laboratory Manual, 3rd ed. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Sanger F, Nicklen S & Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 74:5463–5467.

Smith M (1994) Nobel lecture. Synthetic DNA and biology. *Biosci Rep* 14:51–66.

Southern EM (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J Mol Biol* 98:503–517.

The *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815.

The *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282:2012–2018.

Venter JC, Adams MA, Myers EW et al (2000) The sequence of the human genome. *Science* 291:1304–1351.

### Studying Gene Expression and Function

Boone C, Bussey H & Andrews BJ (2007) Exploring genetic interactions and networks with yeast. *Nature Rev Genet* 8:437–449.

Botstein D, White RL, Skolnick M & Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* 32:314–331.

DeRisi JL, Iyer VR & Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278:680–686.

International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* 437:1299–320.

Lockhart DJ & Winzeler EA (2000) Genomics, gene expression and DNA arrays. *Nature* 405:827–836.

Mello CC & Conte D (2004) Revealing the world of RNA interference. *Nature* 431:338–342.

Nusslein-Volhard C & Weischaus E (1980) Mutations affecting segment number and polarity in *Drosophila*. *Nature* 287:795–801.

Palmiter RD & Brinster RL (1985) Transgenic mice. *Cell* 41:343–345.

Rubin GM & Sprading AC (1982) Genetic transformation of *Drosophila* with transposable element vectors. *Science* 218:348–353.

Sabeti PC, Schaffner SF, Fry B et al (2006) Positive natural selection in the human lineage. *Science* 312:1614–1620.

Weigel D & Glazebrook J (2001) *Arabidopsis:* A Laboratory Manual. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.


LCY Biotechnology Holding, Inc.
Ex. 1008
Page 236 of 240

they prevent incompletely folded proteins from becoming irreversibly aggregated. Both calnexin and calreticulin also promote the association of incompletely folded proteins with another ER chaperone, which binds to cysteines that have not yet formed disulfide bonds.

Calnexin and calreticulin recognize *N*-linked oligosaccharides that contain a single terminal glucose, and therefore bind proteins only after two of the three glucoses on the precursor oligosaccharide have been removed by ER glucosidases. When the third glucose has been removed, the protein dissociates from its chaperone and can leave the ER.

How, then, do calnexin and calreticulin distinguish properly folded from incompletely folded proteins? The answer lies in yet another ER enzyme, a glucosyl transferase that keeps adding a glucose to those oligosaccharides that have lost their last glucose. It adds the glucose, however, only to oligosaccharides that are attached to unfolded proteins. Thus, an unfolded protein undergoes continuous cycles of glucose trimming (by glucosidase) and glucose addition (by glycosyl transferase), maintaining an affinity for calnexin and calreticulin until it has achieved its fully folded state (**Figure 12–53**).

## Improperly Folded Proteins Are Exported from the ER and Degraded in the Cytosol

Despite all the help from chaperones, many protein molecules (more than 80% for some proteins) translocated into the ER fail to achieve their properly folded or oligomeric state. Such proteins are exported from the ER back into the cytosol, where they are degraded. The mechanism of retrotranslocation, also called *dislocation*, is still unknown but is likely to be similar to other post-translational modes of translocation. For example, like translocation into mitochondria or chloroplasts, chaperone proteins are probably necessary to keep the polypeptide chain in an unfolded state prior to and during transport. Similarly, a source of energy is required to provide directionality to the transport and to pull the protein into the cytosol. Finally, a translocator, possibly composed of some of the same components that are used for forward transport into the ER (such as Sec61), is presumably necessary.

Selecting proteins from the ER for degradation is a challenging process. Misfolded proteins or unassembled protein subunits should be degraded, but folding

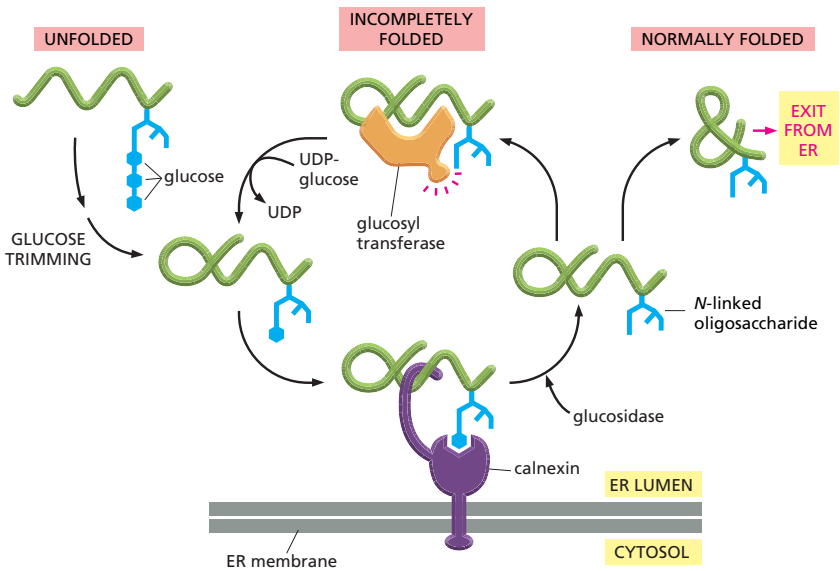



**Figure 12–53 The role of *N*-linked glycosylation in ER protein folding.** The ER-membrane-bound chaperone protein calnexin binds to incompletely folded proteins containing one terminal glucose on *N*-linked oligosaccharides, trapping the protein in the ER. Removal of the terminal glucose by a glucosidase releases the protein from calnexin. A glucosyl transferase is the crucial enzyme that determines whether the protein is folded properly or not: if the protein is still incompletely folded, the enzyme transfers a new glucose from UDP-glucose to the *N*-linked oligosaccharide, renewing the protein's affinity for calnexin and retaining it in the ER. The cycle repeats until the protein has folded completely. Calreticulin functions similarly, except that it is a soluble ER resident protein. Another ER chaperone, ERp57 (not shown), collaborates with calnexin and calreticulin in retaining an incompletely folded protein in the ER.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 237 of 240

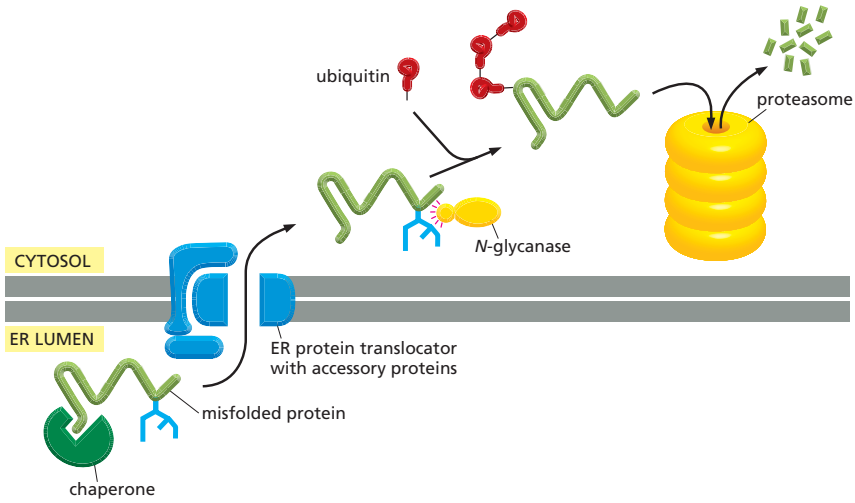



**Figure 12–54 The export and degradation of misfolded ER proteins.** Misfolded soluble proteins in the ER lumen are translocated back into the cytosol, where they are deglycosylated, ubiquitylated, and degraded in proteasomes. Misfolded membrane proteins follow a similar pathway.

intermediates of newly made proteins should not. Help in making this distinction comes from the *N*-linked oligosaccharides, which serve as timers that measure how long a protein has spent in the ER. The slow trimming of a particular mannose on the core-oligosaccharide tree by an enzyme (a mannosidase) in the ER is thought to create a new oligosaccharide structure that the retrotranslocation apparatus recognizes. Proteins that fold and exit from the ER faster than the action of the mannosidase would therefore escape degradation.

Once a misfolded protein has been retrotranslocated into the cytosol, an *N*-glycanase removes its oligosaccharide chains *en bloc*. The deglycosylated polypeptide is rapidly ubiquitylated by ER-bound ubiquitin-conjugating enzymes and is then fed into proteasomes (discussed in Chapter 6), where it is degraded (**Figure 12–54**).

## Misfolded Proteins in the ER Activate an Unfolded Protein Response

Cells carefully monitor the amount of misfolded protein in various compartments. An accumulation of misfolded proteins in the cytosol, for example, triggers a *heat-shock response* (discussed in Chapter 6), which stimulates the transcription of genes encoding cytosolic chaperones that help to refold the proteins. Similarly, an accumulation of misfolded proteins in the ER triggers an **unfolded protein response**, which includes an increased transcription of genes encoding ER chaperones, proteins involved in retrotranslocation and protein degradation in the cytosol, and many other proteins that help to increase the protein folding capacity of the ER.

How do misfolded proteins in the ER signal to the nucleus? There are three parallel pathways that execute the unfolded protein response (**Figure 12–55**A). The first pathway, which was initially discovered in yeast cells, is particularly remarkable. Misfolded proteins in the ER activate a transmembrane protein kinase in the ER, which causes the kinase to oligomerize and phosphorylate itself. (Some cell-surface receptors in the plasma membrane are activated in a similar way, as discussed in Chapter 15). The oligomerization and autophosphorylation activates an endoribonuclease domain in the cytosolic portion of the same molecule, which cleaves a specific, cytosolic RNA molecule at two positions, excising an intron. The separated exons are then joined by an RNA ligase, generating a spliced mRNA, which is translated to produce an active gene regulatory protein. This protein activates the transcription of the genes encoding the proteins that mediate the unfolded protein response (Figure 12–55B).

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 238 of 240

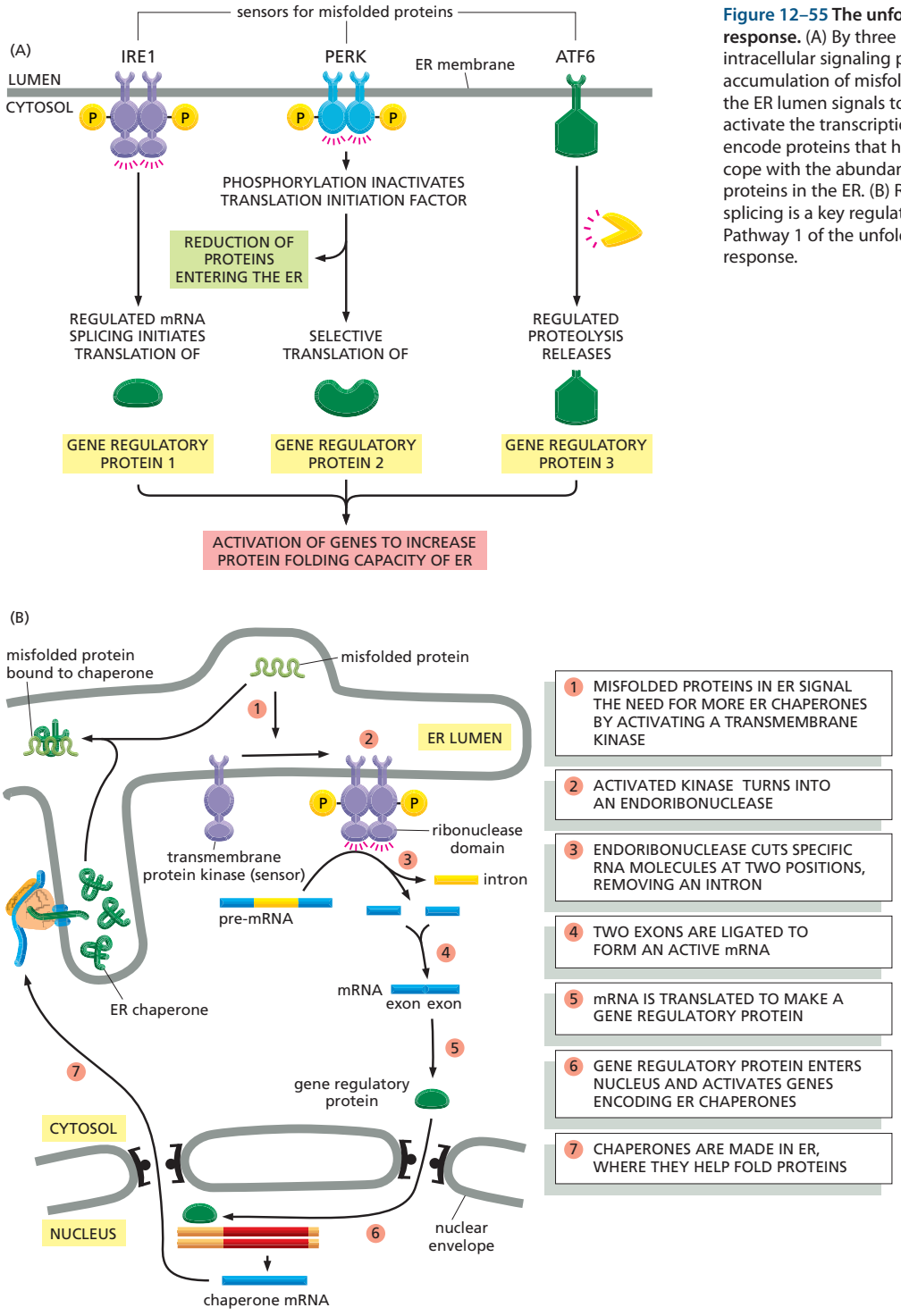



**Figure 12–55 The unfolded protein response.** (A) By three parallel intracellular signaling pathways, the accumulation of misfolded proteins in the ER lumen signals to the nucleus to activate the transcription of genes that encode proteins that help the cell to cope with the abundance of misfolded proteins in the ER. (B) Regulated mRNA splicing is a key regulatory switch in Pathway 1 of the unfolded protein response.

Misfolded proteins also activate a second transmembrane kinase in the ER, which inhibits a translation initiation factor by phosphorylating it and thereby reduces the production of new proteins throughout the cell. One consequence of the reduction in protein translation is to reduce the flux of proteins into the ER, thereby limiting the load of proteins that need to be folded there. Some proteins, however, are preferentially translated when translation initiation factors

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 239 of 240

are scarce (see p. 490), and one of these is a gene regulatory protein that helps activate the transcription of the genes encoding proteins active in the unfolded protein response.

Finally, a third gene regulatory protein is initially synthesized as an integral ER membrane protein. Because it is covalently tethered to the membrane, it cannot activate the transcription of genes in the nucleus. When misfolded proteins accumulate in the ER, the transmembrane protein is transported to the Golgi apparatus, where it encounters proteases that cleave off its cytosolic domain, which can now migrate to the nucleus and help activate the transcription of the genes encoding proteins involved in the unfolded protein response. The relative importance of each of these three pathways differs in different cell types, enabling each cell type to tailor the unfolded protein response to its particular needs.

## Some Membrane Proteins Acquire a Covalently Attached Glycosylphosphatidylinositol (GPI) Anchor

As discussed in Chapter 10, several cytosolic enzymes catalyze the covalent addition of a single fatty acid chain or prenyl group to selected proteins. The attached lipids help to direct these proteins to cell membranes. A related process is catalyzed by ER enzymes, which covalently attach a **glycosylphosphatidyl-inositol (GPI) anchor** to the C-terminus of some membrane proteins destined for the plasma membrane. This linkage forms in the lumen of the ER, where, at the same time, the transmembrane segment of the protein is cleaved off (**Figure 12–56**). A large number of plasma membrane proteins are modified in this way. Since they are attached to the exterior of the plasma membrane only by their GPI anchors, they can in principle be released from cells in soluble form in response to signals that activate a specific phospholipase in the plasma membrane. Trypanosome

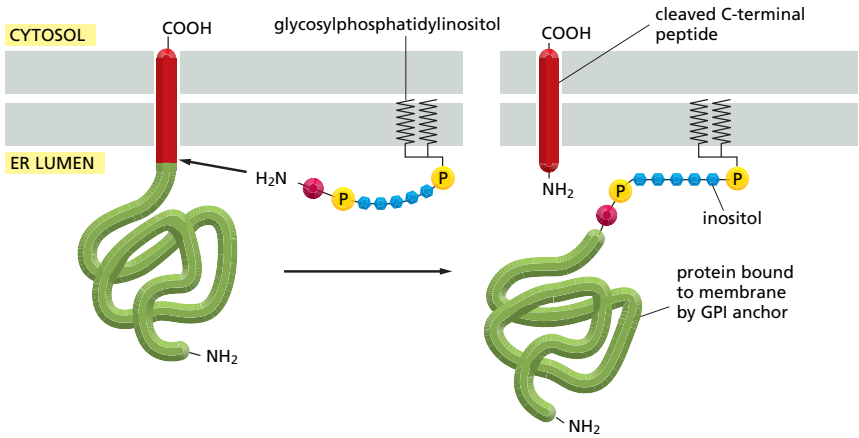



**Figure 12–56 The attachment of a GPI anchor to a protein in the ER.** Immediately after the completion of protein synthesis, the precursor protein remains anchored in the ER membrane by a hydrophobic C-terminal sequence of 15–20 amino acids; the rest of the protein is in the ER lumen. Within less than a minute, an enzyme in the ER cuts the protein free from its membrane-bound C-terminus and simultaneously attaches the new C-terminus to an amino group on a preassembled GPI intermediate. The sugar chain contains an inositol attached to the lipid from which the GPI anchor derives its name. It is followed by a glucosamine and three mannoses. The terminal mannose links to a phosphoethanolamine that provides the amino group to attach the protein. The signal that specifies this modification is contained within the hydrophobic C-terminal sequence and a few amino acids adjacent to it on the lumenal side of the ER membrane; if this signal is added to other proteins, they too become modified in this way. Because of the covalently linked lipid anchor, the protein remains membrane-bound, with all of its amino acids exposed initially on the lumenal side of the ER and eventually on the cell exterior.

LCY Biotechnology Holding, Inc.
Ex. 1008
Page 240 of 240